REPORT RESUMES

ED 003 252 24

EMPIRICAL DETERMINATION OF SAMPLING DISTRIBUTIONS OF ITEM DISCRIMINATION INDICES AND A RELIABILITY COEFFICIENT.

BY- BAKER, FRANK
WISCONSIN UNIV., MADISON, SCH. OF EDUC.

REPORT NUMBER CRP-1299 PUB DATE NOV 62
CONTRACT OEC-2-10-071
EDRS PRICE MF-$0.27 HC-$6.20 155P.

DESCRIPTORS- STATISTICAL STUDIES, *STATISTICAL ANALYSIS, EVALUATION, *ITEM ANALYSIS, MADISON, WISCONSIN

THE RESULTS OF THIS STUDY CLEARLY INDICATE THE INADVISABILITY OF ATTEMPTING TO PERFORM ITEM ANALYSES BASED UPON SMALL SAMPLES. THE SINGLE ITEM DISCRIMINATION INDEX BASED UPON SAMPLES OF SIZE 15, 30, OR 60, HAS LITTLE OR NO VALUE. THOSE BASED UPON SAMPLES OF 120 ARE ONLY INDICATIVE IN NATURE. (LP)

1299

ED 003252

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education
This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

EMPIRICAL DETERMINATION OF SAMPLING DISTRIBUTIONS OF ITEM DISCRIMINATION INDICES AND A RELIABILITY COEFFICIENT

Frank B. Baker

Department of Educational Psychology
School of Education
University of Wisconsin
November, 1962

The research reported herein was performed pursuant to a contract with the United States Office of Education, Department of Health, Education, and Welfare.

ACKNOWLEDGMENTS

The author is indebted to the many persons involved in the present study. Only through the excellent cooperation of Dr. Marvin L. Stein, Director, Numerical Analysis Laboratory, University of Minnesota and Mr. Mike Diffley of his staff, was it possible to debug new computer programs and perform production runs on the distant 1103 computer. Mr. Bill Hollman, Numerical Analysis Laboratory, University of Wisconsin, assisted the author in learning the folkways of fortran programming. The team of graduate and undergraduate assistants, whose diligence and high level of technical competence were indispensable, consisted of Mr. Gene Glass, Miss Julie Joyce, Miss Elizabeth Hopkins, and our superb draftsman, Mr. Hans Sukken.

The entire project was made possible by a grant from the Cooperative Research Branch of the U.S. Office of Education, Department of Health, Education, and Welfare.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS ii

TABLE OF CONTENTS iii

LIST OF TABLES iv

LIST OF FIGURES vi

I. INTRODUCTION 1

  Review of relevant literature 3
  Selected item discrimination indices 5

II. DATA COLLECTION AND PROCESSING PROCEDURES 15

  Selection of the population, the items, criteria, and sample sizes 15
  Data processing procedures 19
  Verification of sampling procedures 23

III. ANALYSIS OF THE EMPIRICAL SAMPLING DATA FOR THE SELECTED ITEM DISCRIMINATION INDICES 26

  Additional symbols 26
  Analysis of the indices 27
  Comparison of sampling properties of the selected indices 82

IV. THE HOYT RELIABILITY COEFFICIENT 87

  Background to the problem 87
  Analysis of the empirical results 91
  Sampling characteristics of the Hoyt reliability coefficient 95

V. SUMMARY AND CONCLUSIONS 97

REFERENCES 106

APPENDIX A 110

APPENDIX B 144

LIST OF TABLES

Table		Page
1	Population values of the item parameters X_{50} and β for the nine selected items, internal criterion	17
2	Population values of the item discrimination indices for the nine selected items based upon internal and external criterion scores	18
3	Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit test of the sample mean for samples of size 15, 30, 60, 120.	23
4	Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of item difficulty for samples of size 15, 30, 60, 120	28
5	Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of biserial r for samples of size 15, 30, 60, 120, internal criterion	32
6	Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of biserial r for samples of size 15, 30, 60, 120, external criterion	38
7	Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of r_f for samples of size 15, 30, 60, 120, internal criterion	42
8	Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of r_f for samples of size 15, 30, 60, 120, external criterion	49
9	Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests for r_c for samples of size 15, 30, 60, 120, internal criterion	53
10	Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of r_c for samples of size 15, 30, 60, 120, external criterion	58
11	Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of delta P for samples of size 15, 30, 60, 120, internal criterion	63
12	Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of delta P for samples of size 15, 30, 60, 120, external criterion	70

Table Page

13 Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of the maximum likelihood estimate of the item parameter X_{50} for samples of size 30, 60, 120, internal criterion 73

14 Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of the maximum likelihood estimate of the item parameter X_{50} of size 30, 60, 120, external criterion 75

15 Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of the maximum likelihood estimate of the item parameter β of size 30, 60, 120, internal criterion 78

16 Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of the maximum likelihood estimate of the item parameter β of size 30, 60, 120, external criterion 80

17 Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of the Hoyt reliability coefficient for samples of size 15, 30, 60, 120 91

LIST OF FIGURES

Figure		Page
1.	Hypothetical item characteristic curve	10
2.	Univac 1103 item analysis program flow chart	20
3.	Frequency distributions of the sample mean obtained from samples of size 15, 30, 60, and 120, internal criterion	24
4.	Theoretical relationship of item difficulty to the item parameters X_{50} and β	30
5.	Observed relationship of item difficulty to the item parameters X_{50} and β based upon samples of size 120	30
6.	Observed relationship of biserial r to the item parameters X_{50} and β as obtained from samples of size 15, internal criterion	34
7.	Observed relationship of biserial r to the item parameters X_{50} and β as obtained from samples of size 120, internal criterion	34
8.	Observed relationship of the standard error of biserial r to the item parameters X_{50} and β as obtained from samples of size 15, internal criterion	34
9.	Observed relationship of the standard error of biserial r to the item parameters X_{50} and β as obtained from samples of size 120, internal criterion	34
10.	Observed relationship of r_f to the item parameters X_{50} and β as obtained from samples of size 15, internal criterion	46
11.	Observed relationship of r_f to the item parameters X_{50} and β as obtained from samples of size 120, internal criterion	46
12.	Observed relationship of the standard error of r_f to the item parameters X_{50} and β as obtained from samples of size 15, internal criterion	46
13.	Observed relationship of the standard error of r_f to the item parameters X_{50} and β as obtained from samples of size 120, internal criterion	46
14.	Observed relationship of r_c to the item parameters X_{50} and β as obtained from samples of size 15, internal criterion	56
15.	Observed relationship of r_c to the item parameters X_{50} and β as obtained from samples of size 120, internal criterion	56

Figure Page

16. Observed relationship of the standard error of r_c to the item parameters X_{50} and β as obtained from samples of size 15, internal criterion 56

17. Observed relationship of the standard error of r_c to the item parameters X_{50} and β as obtained from samples of size 120, internal criterion 56

18. Observed relationship of delta P to the item parameters X_{50} and β as obtained from samples of size 15, internal criterion 66

19. Observed relationship of delta P to the item parameters X_{50} and β as obtained from samples of size 120, internal criterion 66

20. Observed relationship of the standard error of delta P to the item parameters X_{50} and β as obtained from samples of size 15, internal criterion 66

21. Observed relationship of the standard error of delta P to the item parameters X_{50} and β as obtained from samples of size 120, internal criterion 66

22. Frequency distributions of the Hoyt reliability coefficient obtained from samples of size 15, 30, 60, 120 93

23. Cumulative frequency distributions of the Hoyt reliability coefficient and of the product moment correlation coefficient, $\rho = r_H$, $\rho = \sqrt{r_H}$ 94

24. Frequency distribution of item difficulty for the nine items as obtained from random samples of size 15, 30, 60, 120, internal criterion 111

25. Frequency distribution of biserial r for the nine items as obtained from random samples of size 15, 30, 60, 120, internal and external criterion 114

26. Frequency distribution of r_f for the nine items as obtained from random samples of size 15, 30, 60, 120, internal and external criterion 119

27. Frequency distribution of r_c for the nine items as obtained from random samples of size 15, 30, 60, 120, internal and external criterion 124

28. Frequency distribution of delta P for the nine items as obtained from random samples of size 15, 30, 60, 120, internal and external criterion 129

Figure **Page**

29. Frequency distribution of the maximum likelihood estimate of the item parameter X_{50} for the nine items as obtained from random samples of size 30, 60, 120, internal and external criterion 134

30. Frequency distribution of the maximum likelihood estimate of the item parameter β for the nine items as obtained from random samples of size 30, 60, 120, internal and external criterion. 139

31. Project PERT chart 145

INTRODUCTION

Although the role of item discrimination indices is well established in both test construction and psychometric theory, the indices themselves remain somewhat of an enigma. Textbooks on educational measurement often provide a rationale for the application and computation of item discriminating indices, yet, typically do not provide a basis for their interpretation, a description of the conditions which affect them, or list the idiocrasies of the indices. A meaningful interpretation of the numerical value of an item discrimination index can be made only when the sampling properties of the index and the conditions which affect these properties are known. Over a decade ago, Guillikson [19] indicated a need for knowledge of the sampling properties of item discrimination indices; he stated:

> "In addition to systematic changes of item parameters with group ability and with test length, the item statistics are subject to random sampling variation. The magnitude of such fluctuations should be determined in order that we may estimate the change in test parameters to be expected when the test is used on a new sample. These sampling errors could also be used to determine the size for the item analysis that is necessary to give reasonable sampling stability in the test parameters."

The present study is designed to provide empirical evidence as to the magnitude of the fluctuations of item discrimination indices computed from small samples, thus replying in part to the research needs indicated by Guilliksen [19].

The need for information as to the sampling properties of item discrimination indices has existed for many years, yet studies of this type are rare. Theoretical studies are hampered by the complexity of the mathematics involved and by a dearth of previous work upon which to base subsequent developments. In situations of this type, the Monte Carlo technique has proven to be extremely useful. In the past, however, the exhorbitant computational demands of this approach to sampling distributions have precluded any extensive research of this type in education or psychology. The removal of the computational barrier has been accomplished through the development of digital computer programs [2] which perform item analysis. The existing digital computer programs [2] provide present researchers an opportunity to study item discrimination indices in much greater detail than has heretofore been possible. Without the digital computer, the computation of sufficient numbers of item discrimination indices to verify the sampling properties of such indices is economically impractical.

In the present study, the empirical sampling distributions were obtained through application of the Monte Carlo technique to items of a test. In order to maximize the usefulness of the obtained empirical sampling distributions, the following procedures were employed in the study: items having various types of item characteristic curves were selected, internal and external criterion scores were used, and four sample sizes were selected. It was anticipated that the above conditions would provide a somewhat systematic set of data with which the sampling properties of item discrimination indices could be studied. The resulting data will be used to examine the following five questions:

1. What are the effects of sample size upon the sampling properties of the indices?
2. Are the correlational type indices distributed as the product moment correlation coefficient?
3. What are the relationships between the parameters of the item characteristic curve and the sampling properties of the indices?
4. How do the sampling distributions of indices computed with internal and external criterion differ?
5. Under the conditions established, what are the relative merits of the sampling properties of the selected indices?

The exploration of the above topics should aid in providing a better basis for the selection and interpretation of item discrimination indices by both the measurement specialist and the test constructor.

A subsidiary part of the present investigation is concerned with the sampling properties of the reliability coefficient as computed by the Analysis of Variance Method, Hoyt [21]. Although there is a strong technical similarity between this avenue of inquiry and the main body of the study, a separate section will be devoted to the empirical sampling properties of this reliability coefficient.

Relevant Literature

Although a considerable body of literature pertaining to item discrimination indices has been published, little of this literature proved to be relevant to sampling distributions of item discrimination indices. The author was unable to locate a single theoretical paper dealing with this topic. The empirical item analysis studies tended to follow the pattern set by Cook [7], Vernon [43], and the classical work of Long and Sandiford [32] in which many different indices were computed for the same data, lists of the "best" items selected by each method obtained, and the number of items common to the several lists noted. Such studies aid the user in selection of an index, but do not provide a firm basis for establishing a minimum sample size for the technique.

Only the study by Maritz [34] was designed to investigate the sampling properties of an item discrimination index. Maritz [34] studied the empirical sampling distribution of a G statistic which is a function of the standard deviation of the normal ogive fitted to item response data. Rather than using data obtained from actual tests, he drew 100 random samples of 200 cases each from a known bivariate normal distribution having a correlation of .600; the G statistic was then computed for each of the 100 samples. The mean value of the obtained G's was .610, the observed variance of the obtained G's was .0084 and the theoretical variance was .0079. Because of the close agreement of the empirical and theoretical values, Maritz concluded the G statistic was nearly normally distributed for samples of size 200.

Even though the item discrimination index is fundamental to modern mental measurement theory as formulated by Lawley [31] and extended by Lord [33] and others, the dearth of research concerning the sampling distributions of item discrimination indices suggests that an important area has been neglected.

Selected Item Discrimination Indices

Within the large number of available indices, considerable redundancy exists as to type of index, therefore, the indices were classified into three general types and indices representative of each type selected. The three categories and their associated indices are given below.

A. Correlational

  1. Biserial r (r_b)
  2. Product moment correlations estimated by the U-L 27% method. (r_f)

B. Heuristic

  1. Cook's Index B (r_c)
  2. Delta P, the difference in the proportion of persons in upper and lower 27% extreme groups answering the item correctly. (Pd)

C. Curve Fitting

  1. Maximum likelihood estimation of the item parameters X_{50} and β.

Each of these indices is discussed in greater detail below.

Biserial r

If one considers the response to the item to be normally distributed but obtainable only as a dichotomus response, the item-criterion correlation will be the biserial r. The computational formula given below is from Walker and Lev [45] and was implemented in the 1103 computer program.

$$(1)\quad r_b = \frac{\overline{X}_1 - \overline{X}}{S_x}\left(\frac{P}{z}\right)$$

Where: $\overline{X}_1$ is the mean score of all persons answering the item correctly.

$\overline{X}$ is the mean of the sample.

S_x is the standard deviation of the sample.

P is the proportion of persons answering the item correctly.

z is the ordinate of the normal curve at the deviate which divides the area of the unit normal curve into P and (1 - P).

The derivation of biserial r, due to Pearson [36], did not provide the sampling distribution of the index. Johnson [25[indicated the sampling distribution of biserial r has not yet been derived. Assuming large samples, Sopher [41] gave the standard error of r_b as:

$$(2)\quad \sigma_{r_b} = \frac{1}{N}\left\{\frac{PQ}{Z^2} - \left[\frac{3}{2} + \left(1 + \frac{PX}{Z}\right)\left(1 + \frac{QX}{Z}\right)\right] r^2 + r^4\right\}$$

Where: X is the normal deviate corresponding to P, with P Q.

N is the sample size.

The standard error reported by Sopher [41] is unwieldy, hence several approximate formulas have been devised. The approximation due to McNamara and Dunlap [35] is presented in equation 3 and was used in the present study.

$$(3)\quad \sigma_{r_b} = \frac{\frac{\sqrt{PQ}}{Z} - R_b^2}{\sqrt{N}}$$

The lack of a known sampling distribution for biserial r poses the problem of how to evaluate the empirical sampling distribution of biserial r obtained in the study. The product moment correlation is sufficiently different from the biserial correlation so that it serves only as a rough guide. Because of the lack of theoretical framework, the decision made was to report the empirical results, the large sample standard errors of biserial r, the mean r, and σ_r, of the product moment correlation coefficient for the population value of r_b for illustrative but not rigorous comparisons. In addition the Komologorov-Smirnov [40] test was employed to test the goodness of fit of the empirical distribution of r_b against that of the product moment correlation.

Product Moment Correlation Estimated by the U-L 27% Method

The use of the two groups of individuals comprising the upper and lower 27% of the criterion distribution is based upon the optimum group size, 27%, derived by Kelly about 1927, though formal publication of the proof did not occur until 1939. Kelly's [28] proof was the basis for a widely used system of item analysis developed by Flanagan [14]. Flanagan prepared a table based upon Pearson's tables of the bivariate normal correlation surface which yielded the product moment correlation coefficient corresponding to the proportion of correct response in the upper and lower 27% extreme groups.

Fan [11] prepared tables similar in purpose to those of Flanagan [14] which were based upon a more accurate method of interpolation in Pearson's tables of the bivariate normal correlation surface; therefore, Fan's tables are to be preferred.

In the present study, the computer determined the proportion of persons in the extreme 27% groups answering the items correctly. Fan's tables [10] were then used to obtain the values of r_f. The values of r_f were punched into Hollerith cards for subsequent processing.

Because the U-L 27% method estimates the product moment correlation coefficient, the empirical results can be compared directly with the known sampling distribution of the product moment correlation.

Cook's Index B

The heuristic index due to Cook [7], index B, appears in the literature with several different names and methods of derivation; the most recent of which was given by Fowler [15]. The index was computed in the present study using the formulas due to Fowler [15] but shall be designated r_c to indicate its origins.

The rationale used by Fowler [15] in deriving the index is different than the usual item-criterion correlation approach. He considered the index to be the product moment correlation of the answer pattern and the "ideal" answer pattern. The ideal answer pattern was that obtained when perfect discrimination occurred at the score which dichotomized the sample into two groups corresponding to the observed number correct and incorrect answers to the item. Those persons scoring below this point of dichotomy, yet, answering the item correctly, were considered to be missclassified by the item. The formula for the index was obtained by substituting response frequencies into the computation formula of the product moment correlation coefficient. The formula for the index is given in equation 4 below:

$$(4) \quad r_c = \frac{N_1N_2 - NW}{N_1N_2}$$

Where: N_1 is the number of persons answering the item correctly.

N_2 is the number of persons answering the item incorrectly.

W is the number of persons of the lowest N_2 scoring persons who answered the item correctly.

Though the index was derived differently by Cook [7] and Fowler [15], the results were identical; Fowler, however, did not report this relationship. If the index is in fact a product moment correlation coefficient, its sampling properties should be those of the correlation coefficient.

Delta P

The use of the difference in number or proportion of the upper and lower extreme groups answering the item correctly was reported by Cook [7], Long and Sandiford [32], and Johnson [24]. The measurement textbook by Ross and Stanley [39] also reports a variant of this technique. In the present study, the index was computed using the formula given below.

(5) Pd = Pu - Pl

Where: Pu is the proportion of persons in the upper 27% of the criterion variable distribution who answered the item correctly.

Pl is the proportion of the persons in the lower 27% of the criterion variable distribution, who answered the item correctly.

The delta P index is rather interesting as, within a certain range of item difficulty and discrimination, its numerical value closely approximates the numerical value of the correlation coefficient obtained from Fan's tables [10]. The index is not a correlation coefficient and its sampling properties would not be those of the correlation coefficient.

The sampling distribution of Pd can be obtained directly from that of the binomial, hence the empirical sampling distributions can be compared against the latter. In addition, the sampling distribution of the index was compared against the product moment correlation coefficient to ascertain if its properties follow those of the correlation coefficient for any significant range of values.

Maximum Likelihood Estimates of the Item Parameters X_{50} and β

If the proportion of individuals who answer an item correctly at each criterion score is plotted against the criterion variable scale, a smooth line through these points forms a curve which is designated the item characteristic curve.

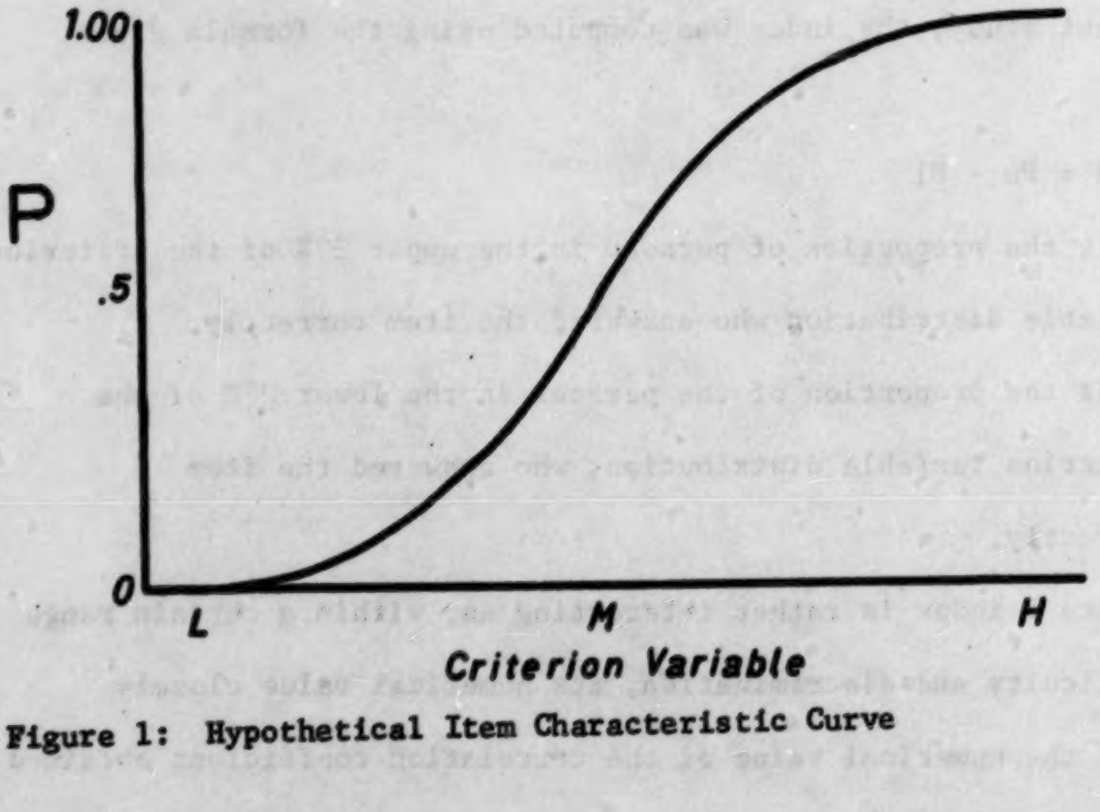


Figure 1: Hypothetical Item Characteristic Curve

The location of the mid-point of the item characteristic curve along the criterion scale indicates the ability level of the subjects at which the item functions. A curve displaced to the left in Figure 1 represents an item functioning among low scoring individuals. Displacement to the right in Figure 1 would represent an item functioning among the high scoring individuals. A long flat curve shows the probability of correct response increases only slightly with increased criterion score, i.e., the item possesses low discrimination. A steep item characteristic curve a small increase in the criterion scale is accompanied by a marked increase in the probability of correct response.

In the present study it was assumed the item characteristic curves were in the form of the integrated normal function, the normal ogive. The parameters of the normal ogive are its mean and standard deviation; however, for purposes of item specification it is convenient to redefine the parameters slightly. The parameters of the item characteristic curve which specify tbe normal ogive fitted to the item response data are the following:

X_{50}, the criterion score at which the probability of correct response is .5. The parameter is expressed in units of the criterion variable standard deviation.

β, a measure of the steepness of the item characteristic curve which specifies the capability of the item to discriminate between the individuals possessing various amounts of the criterion ability. This parameter is the reciprocal of the standard deviation of the fitted normal ogive.

The item parameters obtained from fitting the normal ogive to the item response data hold a fundamental position in mental measurement theory. The classical paper due to Lawley [31] applied the method of maximum likelihood to the estimation of the item parameters. Lawley's paper was followed shortly by a paper by Finney [13] in which estimation of item parameters by maximum likelihood was also proposed. Lawley [31], in addition, showed an individual's true score, test reliability, and other test parameters were functions of the item parameters X_{50} and β.

Fitting of the normal ogive to the observed item response data is accomplished by solving the following equations iteratively until δa and δb are sufficiently small.

$$(6) \quad \delta a\Sigma nw + \delta b\Sigma nwx = \Sigma nwy$$

$$\delta a\Sigma nwx + \delta b\Sigma nwx^2 = \Sigma nwxy$$

or:

$$\delta b_i = \frac{\Sigma nwxy - \frac{\Sigma nwx\ \Sigma nwy}{\Sigma nw}}{\Sigma nwx^2 - \frac{(\Sigma nwx)^2}{\Sigma nwx}}$$

$$\delta a_i = \bar{y} - b\bar{x} \text{ and } \bar{x} = \frac{\Sigma nwx}{\Sigma aw}$$

$$\bar{y} = \frac{\Sigma nwy}{\Sigma nw}$$

Note: All subscripts were omitted in the equations above for purposes of clarity.

$$(7) \quad a_i = a_{i-1} + \delta a_i$$

$$b_i = b_{i-1} + \delta b_i$$

Where: $w = Z^2/PQ$

$y = \frac{p-P}{Z}$

p = observed proportion of individuals passing the item at a given score X.

P = the estimated proportion of individuals passing the item obtained from the integral of the normal function with $a_i + b_i x$ as the entering argument.

Z = normal ordinate corresponding to P

x = the criterion score expressed in standard deviation units.

The derivation of the equations is given in great detail by Garwood [16].

The maximum likelihood estimates of the item parameters X_{50} and β, obtained when δa and δb are sufficiently small, are as follows:

$$(8) \quad \hat{X}_{50} = -a_i/b_i$$

$$\hat{\beta} = b_i$$

The large sample variances of the estimates of X_{50} and β are given by Garwood [16] as:

$$(9) \quad S_b^2 = \frac{1}{\Sigma nw \, (x-\bar{x})^2}$$

$$(10) \quad S_{X_{50}}^2 = \frac{1}{b^2} \left[\frac{1}{(\Sigma nw)^2} + S_b^2 \, (X_{50} - \bar{X})^2 \right]$$

The appropriate small sample variances are not known, Finney [13].

The computer program used to compute the maximum likelihood estimates of X_{50} and β has several characteristics which should be noted. First:

the test scores were not pooled into intervals along the criterion scale, thus, each raw score constituted an interval represented by X_i in the equations above. When small samples are used the observed proportions of response at the X_i points consist chiefly of 1's and 0's. A second characteristic is: the sub-routine for evaluating the normal function had an upper bound of 4σ; consequently when the entering argument $y = a_i + b_i x_i$ exceeded 4σ, standard values for Pi and Zi were obtained. The effects of such a practice are negligible.

In item analysis, the occurrence of items having very low or zero discrimination is not unusual; such items yield unstable values for X_{50}. The discrimination index β appears in the denominator of the formula for X_{50}, thus, as β approaches zero, X_{50} approaches infinity. The products of this situation are extremely large values of X_{50} which are of no practical value.

The item parameters X_{50} and β are independent and normally distributed, hence the empirical sampling distributions can be compared with the normal distribution.

DATA COLLECTION AND PROCESSING PROCEDURES

Selection of the Population, Items, Criteria, and Sample Sizes

The basic data of the present study consisted of 747 test papers drawn from a pool of approximately 6,000 ACE-Q tests administered as part of a university admissions program. Assuming normality, the mean (37.99) and standard deviation (10.40) of the local norms were used to compute the number of subjects, in a group of 750 persons, expected to obtain each of the raw scores within the range of the instrument. The 6,000 test papers were then stratified by raw score and the expected number of subjects at a given raw score were drawn at random from those available at the given score. The random selection process was continued until a normally distributed sample of 747 (due to rounding of expected frequencies) test papers was obtained. The group of 747 test papers will be referred to as the population throughout the present study. The first semester freshman grade point averages of the 747 subjects were obtained to serve as the external criterion in the item analysis. The distribution of the grade point averages was normalized by means of an area transformation employing the mean and standard deviation of the local norms. The above procedure insured that the internal and external criterion scores were normally distributed with the same mean and standard deviation.

The item response choices made by the 747 subjects were punched into Hollerith cards in order to read the test papers into computer memory. In addition, the test scores and the normalized grade point averages were punched into cards for computer input.

Because of considerable redundancy of item characteristics within the instrument, it was not necessary to utilize all items in the sampling investigation. It was felt desirable to select representative items from the domain of item parameter pairs; thus reducing the computing load at a minimum loss of information. Three levels of X_{50}, high, medium, and low, were selected in order to have items which function at these segments of the criterion scale. In order to have various amounts of discrimination at each level of X_{50}, three levels of β, high, medium, and low, were chosen. The population values of the item parameters X_{50} and β, based upon an internal criterion, were inspected to locate items which approximated the nine pairs of item parameter values needed. If the data were contrived, one could establish "nice" values within the item parameter pairs, however, the empirical data necessitated using the best of the available pairs. The nine items selected and their item parameters are presented in Table 1.

The same items were used throughout the item analysis even though the item parameter pairs were chosen only on the basis of the internal criterion. The deliberate lack of order to the item parameters based upon the external criterion had two purposes: one, to emphasize the shift in parameter value which occurs when the criterion is changed; and two, to determine if the resulting sampling distributions merely follow that of a smaller value of the item discrimination index or are affected in an amount greater than that due to the

Table 1: Population values of the item parameters X_{50} and β for the nine selected items, internal criterion.

Item*	Level		Parameter Value	
	X_{50}	β	X_{50}	β
5	High	Low	1.868	.239
9	High	Medium	1.169	.636
6	High	High	1.290	.714
2	Medium	Low	-.050	.358
3	Medium	Medium	.011	.610
8	Medium	High	-.136	.929
7	Low	Low	-1.204	.360
1	Low	Medium	-1.068	.532
4	Low	High	-1.753	.954

* The item numbers used refer to their use in the present study, not the item number in the A. C. E. test.

changed index value. Table 2 presents the population values of the indices employed in the present study based upon internal and external criterion scores. Although the population has only 747 cases, it was assumed the obtained index values were the true values of the indices.

One objective of the present study was to determine the effect of sample size upon the sampling characteristics of the item discrimination indices, with particular reference to small samples. The following

Table 2. Population values of the item discrimination indices for the nine selected items based upon internal and external criterion scores.

	P	r_{bis}		Fowler		r_{Fan}		Delta P		X_{50}		β	
	I	I	E	I	E	I	E	I	E	I	E	I	E
1	.692	.470	.197	.259	.102	.500	.200	.411	.172	-1.068	-2.506	.532	.205
2	.506	.337	.200	.202	.106	.340	.170	.342	.172	- .050	- .077	.358	.205
3	.497	.524	.228	.347	.170	.520	.230	.520	.227	.011	.032	.610	.234
4	.888	.711	.189	.463	.034	.630	.180	.312	.074	-1.753	-6.119	.954	.202
5	.332	.231	-.013	.119	-.050	.230	-.040	.198	-.020	1.868	-34.555	.239	-.013
6	.229	.580	.068	.363	.044	.630	.000	.451	.015	1.290	11.050	.714	.067
7	.659	.340	.186	.208	.071	.300	.200	.282	.173	-1.204	-2.170	.360	.192
8	.534	.695	.142	.532	.091	.670	.120	.683	.120	- .136	- .603	.929	.144
9	.265	.550	.149	.347	.100	.550	.110	.450	.090	1.169	4.266	.636	.149

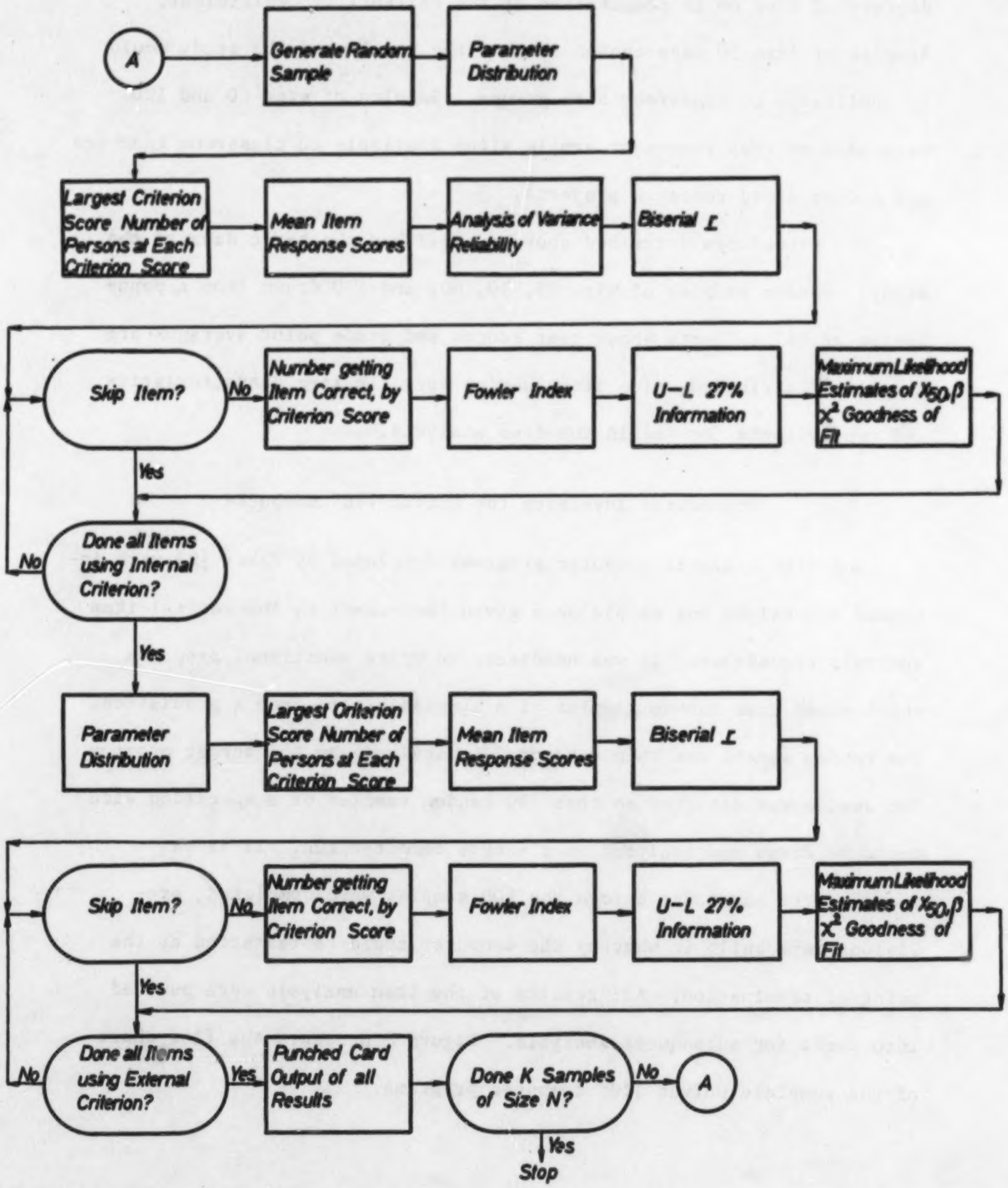
sample sizes were chosen: 15, 30, 60, and 120. Samples of size 15 were chosen to obtain the effect of a very small sample and to reduce the degrees of freedom in computation of the reliability coefficient. Samples of size 30 were chosen so that the results of the study would be applicable to classroom size groups. Samples of size 60 and 120 were used as they represent sample sizes available to classroom teachers and modest sized research projects.

The procedures described above established the basic data of the study: random samples of size 15, 30, 60, and 120 drawn from a population of 747 subjects whose test scores and grade point averages are normally distributed, nine items having specific item characteristics, and two criteria for use in the item analysis.

Procedures Involving the Univac 1103 Computer

The item analysis computer programs developed by Baker [2] were intended to analyze one sample on a given instrument by the several item analysis techniques. It was necessary to write additional programs which would draw random samples of a specified size from a population. The random sample was then subjected to analysis by the former program. The system was designed so that 100 random samples of a specified size could be drawn and analyzed in a single computer run. If it was necessary to terminate before the 100 samples were completed, provisions were built in whereby the computer could be restarted at the point of termination. All results of the item analysis were punched into cards for subsequent analysis. Figure 2 presents the flow chart of the complete Univac 1103 computer programs.

Figure 2. UNIVAC 1103 item analysis program flow chart.


A
Generate Random Sample
Parameter Distribution
Largest Criterion Score Number of Persons at Each Criterion Score
Mean Item Response Scores
Analysis of Variance Reliability
Biserial r
Skip Item?
No
Number getting Item Correct, by Criterion Score
Fowler Index
U-L 27% Information
Maximum Likelihood Estimates of X_{50}, β χ^2 Goodness of Fit
Yes
Done all Items using Internal Criterion?
No
Yes
Parameter Distribution
Largest Criterion Score Number of Persons at Each Criterion Score
Mean Item Response Scores
Biserial r
Skip Item?
No
Number getting Item Correct, by Criterion Score
Fowler Index
U-L 27% Information
Maximum Likelihood Estimates of X_{50}, β χ^2 Goodness of Fit
Yes
Done all Items using External Criterion?
No
Yes
Punched Card Output of all Results
Done K Samples of Size N?
No
A
Yes
Stop

Samples of size 15 were used as the pilot run in order to check the computer programs and subsequent analysis procedures. Due to the excessive number of times the limits of the program were exceeded, the estimation of X_{50} and β was not performed for samples of size 15. Following the initial analysis of samples of 15, 100 samples of size 30, 60, and 120, were analyzed.

The initial production runs indicated more than half of all computer time was devoted to the maximum likelihood estimation of X_{50} and β. In order to obtain a larger number of samples, within the funds available, it was decided to eliminate this technique in the subsequent analyses. The savings in computer time permitted the computation of an additional one hundred samples of size 15, 30, 60, and 120, the end result being 100 samples of X_{50} and β and 200 samples of all other indices for each of the four sample sizes.

Procedures Involving the Control Data 1604 Computer

The results of the 1103 computer analysis were subjected to further analysis using computer programs written for the 1604 computer. The first of the 1604 programs (MOVAR) computed the first four moments, the standard deviation, $\sqrt{\beta_1}$, and β_2 of the obtained item discrimination indices for each criterion and sample size. The second 1604 program (PLOT) compiled the frequency distribution and the cumulative frequency distribution of the obtained indices for each sample size and criterion. The results of the PLOT program were used to prepare graphs of the sampling distributions and the cumulative frequencies were used in the goodness of fit tests. When all production runs on the 1103 were completed, the resulting data were written on magnetic tape and processed through the MOVAR and PLOT programs on the 1604.

A third 1604 program (FAN) used Pu and Pl to look up the value of the corresponding product moment correlation coefficient in Fan's tables, which were stored in computer memory. The results of the FAN program were punched into cards and subsequently processed by the MOVAR and PLOT programs.

Two additional special purpose programs were written for the 1604 computer. David [8] had prepared tables of the probability integral of the correlation coefficient for samples of size 3 to 400 and various values of rho. David's tables, however, were not sufficiently detailed for purposes of the present study. Computer programs were written which implemented the equations used by David [8] in computing her tables, thus allowing one to compute the probability integral for any value of rho and sample size. The population values of r_b, r_c, r_f, and Pd were used to compute the corresponding distributions of r by means of this program.

Pearson [37] provided tables of $\bar{r}$ and σ_r of the sampling distribution of r for several sample sizes and values of rho. The tables, however, were not sufficiently detailed for purposes of the present study. The formulas used to compute the tables were programmed for the 1604, thus allowing the computation of the theoretical sampling statistics of $\bar{r}$ and σ_r for any desired ρ and N.

The combination of the 1103 generation of the sample indices and the 1604 data analysis programs provided a very smoothly functioning system. When used in conjunction with the specialized programs, the system performed the majority of the data processing and analysis of the present study.

Verification of Sampling Procedures

The empirical data generated by the analysis of the randomly drawn samples contains information which can be used to verify the adequacy of the sampling procedures. When an internal criterion score was used, the Hoyt reliability sub-routine computed the sample mean. If the sampling process was random, the sample means should be normally distributed. Subsidiary checks are provided by the bias, population value less the mean of the obtained values, and the standard deviation of the sample means. Table 3 presents the sampling statistics of the sample means. Figure 3 contains the empirical sampling distribution of the sample means for the four sample sizes.

Table 3: Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of the sample mean for samples of size 15, 30, 60, 120. $\mu = 37.99$, $\sigma = 10.40$.

Sample Size	$\bar{X}$	Bias	$S_{\bar{x}}$	$\sigma_{\bar{x}}$	$\bar{S}_x$	Goodness of Fit D_n
15	38.478	-.488	2.591	2.779	10.301	-
30	38.034	-.044	1.821	1.890	10.218	-
60	38.013	-.023	1.303	1.342	10.453	-
120	37.851	.139	.971	.949	10.328	-

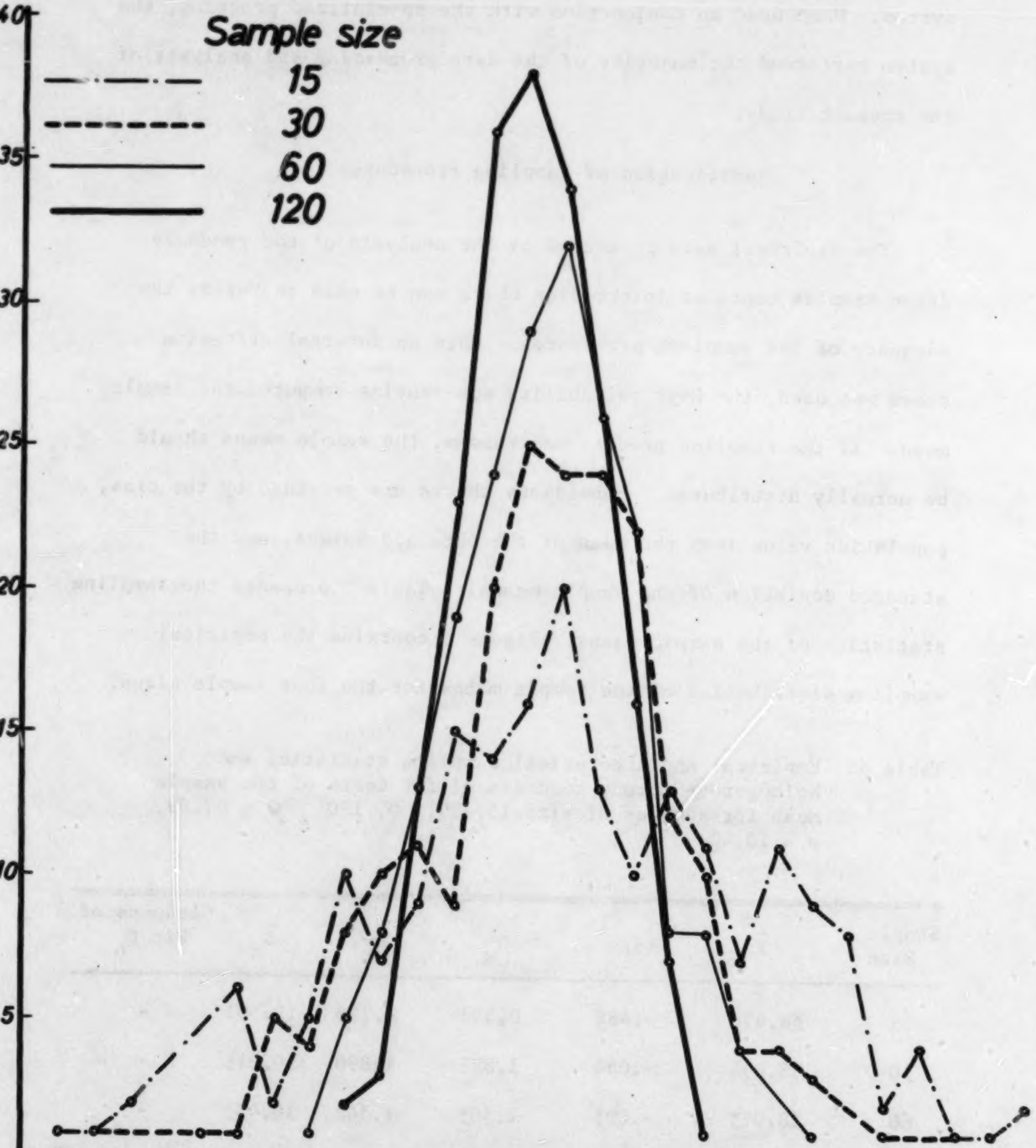


Figure 3. Frequency distributions of the sample mean obtained from samples of size 15, 30, 60, 120, internal criterion.

The biases of all four sample sizes were very small relative to the population means, indicating the mean of the distribution of sample means was a very good estimate of the population mean. In addition, x^2 tests indicated the standard deviations of the sample means, $S\bar{x}$, did not differ significantly from their expected values, $\sigma_{\bar{x}}$. The Kolmogorov-Smirnov goodness of fit tests showed the empirical distribution of the sample means did not differ significantly from the hypothesized normal distribution.

The sampling properties of the sample means indicate the process by which the samples were drawn from the population yields statistics consistent with random sampling procedures.

ANALYSIS OF THE EMPIRICAL SAMPLING DATA FOR THE SELECTED ITEM DISCRIMINATION INDICES

Additional Symbols

The discussion of the sampling characteristics of the item discrimination indices follow the categorization of indices given above. Some additional notation was necessary and is defined below.

$\bar{r}_i$	is the mean value of an index obtained from 200 samples of a given size.
S_{r_i}	is the standard deviation of the 200 index values obtained from samples of a given size.
Bias	is the population value of the index (r_i) less the mean value of the index obtained from the random samples: bias = $\rho_i - \bar{r}_i$.
$\bar{r}$	is the mean value of the product moment correlation coefficient as obtained from the sampling distribution of r for a given sample size.
σ_r	is the standard deviation of the product moment correlation coefficient obtained from the sampling distribution of r for a given sample size.
D_r	indicates the Kolmogorov-Smirnov goodness of fit tests were performed using the cumulative distribution of the product moment correlation coefficient.
D_B	indicates the Kolmogorov-Smirnov goodness of fit tests were performed using the cumulative distribution of the binomial distribution.

D_n indicates the Kolmogorov-Smirnov goodness of fit tests were performed using the cumulative distribution of the normal function.

- no significant difference

* significant difference at .05 level

** significant difference at .01 level

In figures 6 through 21 the relationship between the index and the item parameters X_{50} and β was obtained by plotting the mean index value as a function of the population values of X_{50} and β. The resulting three dimension surface was obtained and smoothed by graphic techniques. The figures depicting the relationships of the standard error of the index to the item parameters X_{50} and β were obtained in a similar fashion. On the surface in each of the figures, three lines were drawn representing high, medium, and low values of β to serve as references.

Analysis of the Indices

Item Difficulty

The classical method of item analysis employs the item-criterion correlation and item difficulty to describe the characteristics of an item, therefore, the empirical sampling distribution of item difficulty was obtained for the nine items. The observed sampling statistics of item difficulty are presented in table 4, accompanied by the bias, σ_p, and the Kolmogorov-Smirnov goodness of fit tests. The bias of the mean item difficulty was small, less than .034, for all items and sample sizes. There did not appear to be a discernible relationship

Table 4: **Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of item difficulty for samples of size 15, 30, 60, 120.**

Item	Sample Size	Observed			Theoretical		Goodness of fit
		$\bar{p}$	Bias	S_p	P	σ_p	D_b
1	15	.682	.010	.119	.692	.119	**
	30	.664	.028	.091	.692	.084	**
	60	.669	.023	.060	.692	.060	**
	120	.662	.031	.041	.692	.042	**
2	15	.540	-.034	.129	.506	.129	**
	30	.508	-.002	.088	.506	.091	-
	60	.510	-.004	.062	.506	.064	**
	120	.517	-.009	.046	.506	.046	-
3	15	.518	-.021	.130	.497	.129	-
	30	.516	-.019	.087	.497	.091	-
	60	.512	-.015	.072	.497	.064	*
	120	.502	-.005	.051	.497	.046	-
4	15	.891	-.003	.079	.888	.081	-
	30	.883	.005	.055	.888	.058	-
	60	.883	.005	.041	.888	.041	-
	120	.885	.003	.028	.888	.029	-
5	15	.327	.005	.132	.332	.122	-
	30	.340	-.008	.084	.332	.086	
	60	.332	.000	.064	.332	.061	-
	120	.335	-.003	.043	.332	.043	-
6	15	.232	-.003	.101	.229	.108	-
	30	.246	-.017	.078	.229	.077	-
	60	.239	-.010	.054	.229	.054	*
	120	.234	-.005	.039	.229	.038	-
7	15	.672	-.013	.121	.659	.122	-
	30	.649	.010	.083	.659	.087	-
	60	.650	.009	.061	.659	.061	-
	120	.652	.007	.047	.659	.043	-
8	15	.548	-.014	.142	.534	.129	-
	30	.551	-.017	.092	.534	.091	**
	60	.543	-.009	.067	.534	.064	-
	120	.543	-.009	.046	.534	.046	**
9	15	.266	-.001	.113	.265	.114	-
	30	.262	-.003	.080	.265	.081	-
	60	.270	-.005	.062	.265	.057	-
	120	.261	-.004	.042	.265	.040	-

between the bias and population value or sample size. The biases had a range of from -.001 to -.034,-.002 to .028, .000 to .023, and .003 to .031 for samples of size 15, 30, 60, and 120 respectively. Only item 1 has biases which were consistently large compared to those yielded by the remaining eight items.

The observed standard deviation closely followed the values of the theoretical standard deviation obtained from the binomial distribution. A high proportion of the observed standard deviations were within .001 of the theoretical value. The largest discrepancy, .013, was yielded by item 8 when samples of 15 were used.

The goodness of fit of the empirical distributions of item difficulty to the corresponding distributions obtained from tables of the cumulative binomial [44] were tested by means of the Kolmogorov-Smirnov goodness of fit test. Twenty six of the thirty six goodness of fit tests indicated the hypothesis was not rejected. When samples of 120 were used, seven of the nine observed distributions followed the theoretical; when samples of size 60, 30, and 15 were used; five of nine, seven of nine, and seven of nine distributions respectively were "good" fits to the theoretical.

An interesting aspect of item difficulty is its relationship to the item parameters X_{50} and β. Item difficulty corresponds to the area under the item characteristic curve and is hence a function of both X_{50} and β. The theoretical relationships for positive β and values of X_{50} from -2σ to $+2\sigma$ are presented in figure 4 and those observed in samples of 120 are presented in figure 5. In figure 4 several lines of equal item difficulty were drawn to indicate the curvilinear nature of

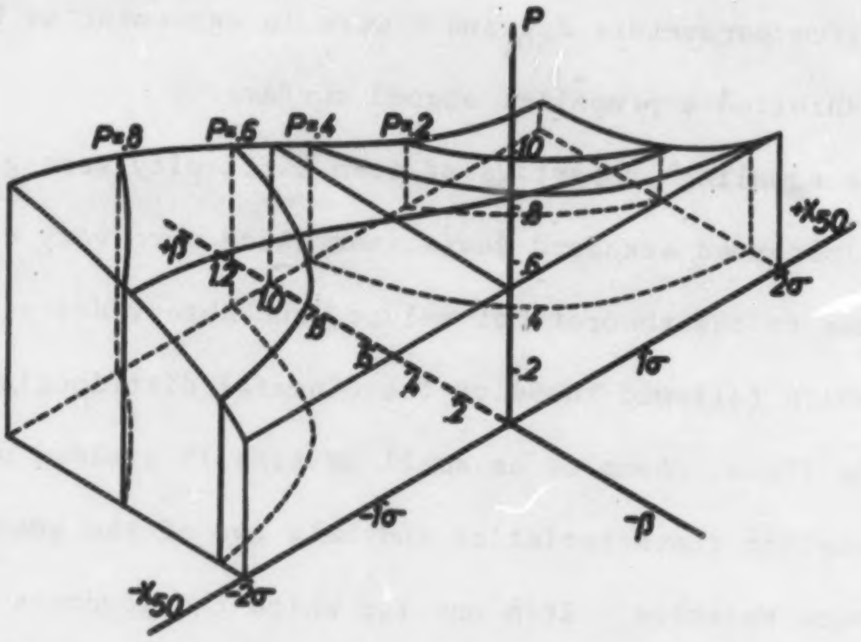


Figure 4. Theoretical relationship of item difficulty to the item parameters X_{50} and β.

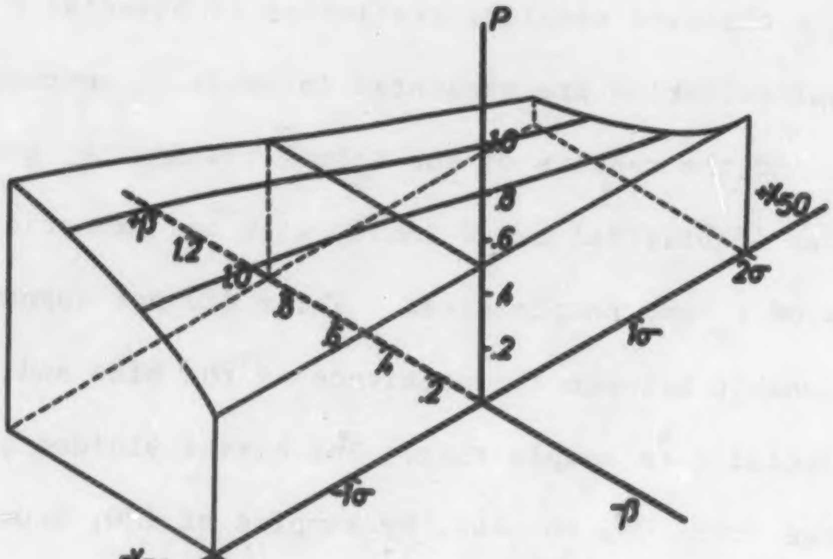


Figure 5. Observed relationship of item difficulty to the item parameters X_{50} and β based upon samples of size 120.

the relationship. The empirical and theoretical relationship of P to the item parameters X_{50} and β were in agreement as both figures 4 and 5 exhibited a propeller shaped surface.

The sampling properties of item difficulty were good: small biases, observed standard deviations which were very close in magnitude to the theoretical values, and observed frequency distributions which followed those of the binomial distribution for eight of the nine items. Samples as small as size 15 yielded uniformly desirable sampling characteristics and only two of the goodness of fit tests were rejected. Item one for which the goodness of fit test lead to rejection four sample sizes, possessed the largest bias of the nine items which might account for the lack of "fit".

The general agreement between the observed sampling characteristics of item difficulty and those of the binomial model indicated that some confidence can be placed in the data sampling procedures.

Biserial r

The observed sampling statistics of biserial r based upon an internal criterion are presented in table 5, accompanied by the bias, $\bar{r}$, σ_r, and the results of the Kolmogorov-Smirnov goodness of fit tests. The bias of biserial r was small, with one exception, for all population values of r_b and sample sizes. There did not appear to be a discernable relationship between the magnitude of the bias and the population value of biserial r or sample size. The biases yielded by samples of size 15 range from .003 to .101, by samples of 120, from .002 to .038. Only two of the nine items, 4 and 5, exhibited a consistent decrease in bias with increased sample size.

Table 5: Empirical and theoretical sampling statistics and Kolmogorov-Smirov goodness of fit tests of biserial r for samples of size 15, 30, 60, 120, internal criterion.

Item	Sample Size	Observed					Theoretical			Large Sam-ple	Goodness of Fit
		r_b	Bias	S_{r_b}	$\sqrt{\beta_1}$	β_2		$\bar{r}$	σ_r	σ_{r_b}	D_r
1	15	.467	.003	.280	.622	3.80	.470	.456	.217	.282	-
	30	.448	.022	.192	.312	3.07	.470	.463	.148	.199	**
	60	.464	.006	.133	.249	3.33	.470	.467	.102	.141	-
	120	.472	-.002	.102	.121	3.26	.470	.469	.069	.100	*
2	15	.359	-.022	.290	.104	2.74	.337	.327	.242	.294	*
	30	.374	-.037	.207	.557	4.03	.337	.332	.166	.208	**
	60	.384	-.047	.158	.228	2.96	.337	.335	.116	.147	**
	120	.364	-.027	.096	.144	2.62	.337	.336	.080	.104	**
3	15	.465	.059	.282	.379	3.10	.524	.510	.204	.253	**
	30	.489	.035	.192	.484	2.82	.524	.517	.138	.179	*
	60	.491	.033	.137	.287	3.05	.524	.521	.096	.126	**
	120	.486	.038	.096	.152	2.80	.524	.523	.063	.089	**
4	15	.610	.101	.339	.689	3.87	.711	.698	.146	.297	**
	30	.674	.037	.249	.839	4.16	.711	.705	.096	.210	**
	60	.718	-.007	.172	.495	3.16	.711	.708	.066	.149	**
	120	.721	-.010	.126	.068	2.90	.711	.711	.034	.105	**
5	15	.286	-.055	.341	.524	3.31	.231	.223	.255	.321	**
	30	.253	-.022	.239	.422	2.74	.231	.227	.177	.227	**
	60	.261	-.033	.108	.048	2.88	.231	.229	.124	.161	**
	120	.264	-.033	.108	.172	3.29	.231	.230	.086	.114	**
6	15	.566	.014	.297	.474	3.87	.580	.566	.189	.271	**
	30	.578	.002	.177	.360	3.14	.580	.573	.127	.192	*
	60	.567	.020	.134	.697	3.92	.580	.577	.088	.136	*
	120	.566	.014	.088	.308	3.01	.580	.579	.056	.096	**
7	15	.320	.020	.351	.273	2.74	.340	.329	.241	.304	*
	30	.300	.040	.212	.155	2.95	.340	.335	.166	.215	**
	60	.332	.008	.[illegible]59	.068	3.02	.340	.337	.116	.152	-
	120	.329	.011	.115	.248	2.93	.340	.339	.080	.107	**
8	15	.683	.012	.235	.740	3.61	.695	.681	.152	.199	**
	30	.703	-.008	.155	.432	2.84	.695	.688	.101	.141	**
	60	.701	-.006	.105	.640	4.02	.695	.692	.069	.100	**
	120	.690	.005	.080	.450	3.30	.695	.694	.038	.071	**
9	15	.545	.005	.305	.623	3.60	.550	.536	.197	.270	**
	30	.559	-.009	.208	.358	2.73	.550	.543	.133	.191	**
	60	.556	-.006	.149	.339	3.25	.550	.547	.092	.135	**
	120	.542	.008	.101	.429	3.01	.550	.549	.060	.095	**

The observed standard deviation of biserial r was compared with its large sample theory value and with that of the product moment correlation coefficient. The differences between the theoretical large sample values and the observed standard deviation of biserial r were small, less than .05, for all sample sizes. When samples of size 15 were used, two of the nine observed standard deviations were within .01 of the large sample values, six of nine, seven of nine, and seven of nine were within .01 of the large sample values when samples of size 30, 60, 120, respectively, were used.

General agreement between the mean value of biserial r and the mean of the product moment correlation coefficient existed; but the agreement was not entirely consistent since it varied with sample size and/or the population value of biserial r. The differences within corresponding pairs of mean values, however, were less than .06. The observed standard deviations of biserial r were consistently larger than the corresponding standard deviations of the product moment correlation coefficient.

The three dimensional relationships among r biserial and the item parameters X_{50} and β are depicted in figures 6 and 7 for samples of size 15 and 120 respectively. If an item discrimination index possesses desirable characteristics its sampling properties should be unaffected by the criterion score level at which the item functions. Thus, as item discrimination increases the three dimensional surface should increase uniformly for all values of X_{50}. Inspection of figure 6 for samples of size 15 shows that biserial r has a slight tendency to decrease in discrimination with increased absolute values of X_{50}. There

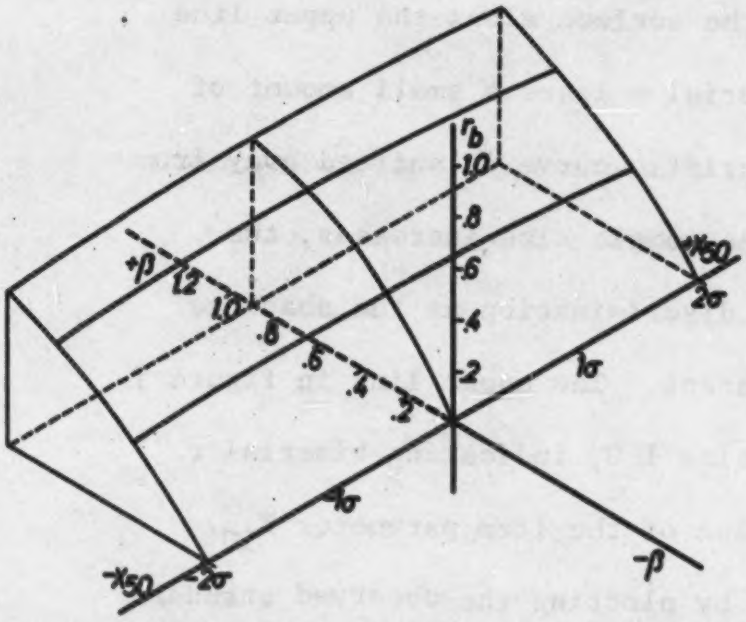


Figure 6. Observed relationship of biserial r to the item parameters X_{50} and β as obtained from samples of size 15, internal criterion.

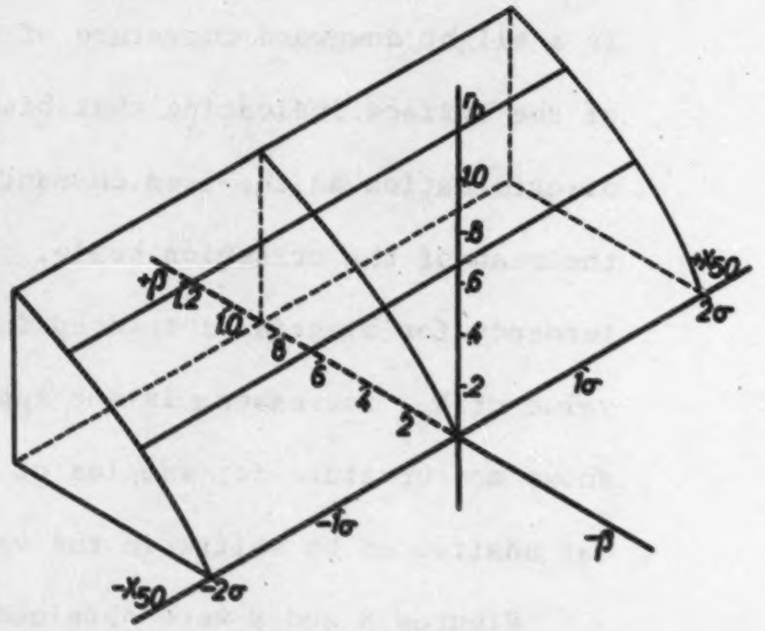


Figure 7. Observed relationship of biserial r to the item parameter X_{50} and β as obtained from samples of size 120, internal criterion.

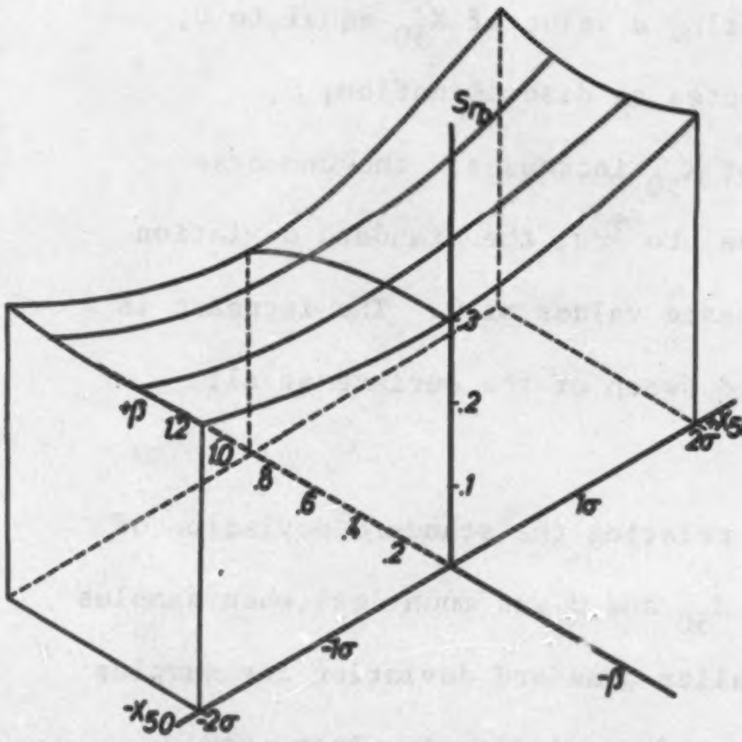


Figure 8. Observed relationship of the standard error of biserial r to the item parameters X_{50} and β as obtained from samples of size 15, internal criterion.

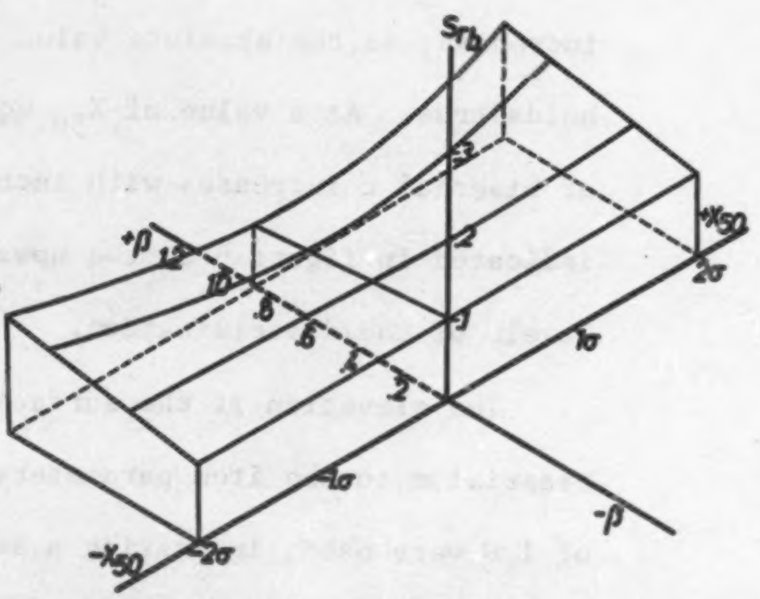


Figure 9. Observed relationship of the standard error of biserial r to the item parameters X_{50} and β as obtained from samples of size 120, internal criterion.

is a slight downward curvature of the surface along the upper line of the surface indicating that biserial r loses a small amount of discrimination as the item characteristic curve is shifted away from the mean of the criterion scale. As sample size increases, the tendency for biserial r to drop in discrimination as the absolute value of X_{50} increases, is not apparent. The upper line in figure 7 shows no curvature for samples of size 120, indicating biserial r was unaffected by shifts in the value of the item parameter X_{50}.

Figures 8 and 9 were obtained by plotting the observed standard deviation as a function of the population values of X_{50} and β. Curves were then fitted through the points by graphic techniques. When samples of size 15 were used, the standard deviation of r_b was relatively large, as indicated by the elevation of the surface in figure 8. Along the line representing a value of X_{50} equal to 0, the standard deviation of r_b decreases as discrimination, β, increases; as the absolute value of X_{50} increases, the converse holds true. At a value of X_{50} equal to $\pm 2\sigma$, the standard deviation of biserial r increases with increased values of β. The increase is indicated in figure 8 by the upward sweep of the surface at all levels of the discrimination.

The elevation of the surface relating the standard deviation of biserial r to the item parameters X_{50} and β was much less when samples of 120 were used, indicating a smaller standard deviation for samples of this size. The upswing of the surface, indicating increased variability with increased magnitude of X_{50} and β, was also much less pronounced than was found for samples of size 15. The formula for the large sample variance of biserial r gives an indication of why this phenomenon was noted.

$$(11) \quad \sigma_{r_b} = \frac{\frac{\sqrt{PQ}}{Z} - r_b^2}{\sqrt{N}}$$

In the numerator of the equation the term $\frac{\sqrt{PQ}}{Z}$ appears, as the ratio becomes larger, the standard deviation increases. The standard deviation of biserial r will decrease when X_{50} equals 0; as $\frac{\sqrt{PQ}}{Z}$ is a constant and the standard deviation of biserial r becomes a function of the subtractive term r_b^2 in the numerator. Hence, as r_b increases the standard deviation decreases. This relationship is shown by the line above the β axis both in figures 8 and 9. Consequently, the standard deviation of biserial r is a function of the location of the item characteristic curve, indicated by X_{50}, and the amount of discrimination indicated by β. One can conclude from figures 6 to 9 that the mean value of biserial r is unaffected by the location of the item characteristic curve along the criterion scale, when samples of 120 were used, and is only slightly impaired when samples as small as 15 were employed. Concurrently, however, the sampling variability of biserial r increases with increased values of the item parameters and/or a reduction in sample size.

External Criterion

The population values of biserial r based upon internal and external criterion were given in table 2. An extremely large drop in the value of biserial r occurred for all items when the criterion variable was switched from total test score to the grade point average. Those items having a high initial value of biserial r exhibited the greatest loss in discriminating power. If the sampling characteristics of an item discrimination index were unaffected by the criterion employed, the sampling statistics based upon the external criterion analysis should merely be those associated with reduced values of the index. The sampling statistics of biserial r based upon an external criterion are given in table 6, accompanied by the bias, $\bar{r}$, σ_r, and the results of the Kolmogorov-Smirnov goodness of fit tests. The bias of biserial r was less than .05 for all population values of biserial r and sample sizes. There did not appear to be a discernible relationship between the magnitude of the bias and the population value or sample size. The biases yielded by the samples of size 15 and 120 ranged in magnitude from .001 to .049. Only two of the items, five, and eight, exhibited a consistent decrease of bias with increased sample size. The biases of three of the items, one, six, and eight, were larger when an external criterion was employed, although the population value was considerably less than that obtained from an internal criterion. The biases obtained from biserial r using an internal and external criterion show very similar behaviors, both having small values which fluctuate inconsistently with sample size and population value of biserial r.

The observed standard deviation of biserial r based upon an external criterion exhibited a reduction in magnitude with increased sample size similar to that of the theoretical large sample standard deviation. The differences between the observed and large sample standard deviations were less than .02. When samples of 120 were used,

Table 6: Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of biserial r for samples of size 15, 30, 60, 120, external criterion.

Item	Sample Size	Observed					Theoretical			Large Sample	Goodness of Fit
		r_b	Bias	S_{r_b}	$\sqrt{\beta_1}$	β_2		$\bar{r}$	σ_r	σ_{r_b}	D_r
1	15	.216	-.019	.333	.033	2.90	.197	.190	.259	.329	*
	30	.221	-.024	.239	.427	3.07	.197	.194	.179	.233	**
	60	.218	-.021	.152	.225	3.25	.197	.195	.125	.164	-
	120	.221	-.024	.106	.317	3.20	.197	.196	.088	.116	**
2	15	.205	-.005	.324	.069	2.61	.200	.193	.258	.313	*
	30	.188	.012	.240	.490	2.76	.200	.197	.179	.222	*
	60	.177	.023	.154	.348	2.92	.200	.198	.125	.157	-
	120	.192	.008	.102	.109	3.13	.200	.199	.088	.111	-
3	15	.226	.002	.332	.322	2.45	.228	.220	.256	.310	**
	30	.234	-.006	.226	.379	3.14	.228	.224	.177	.219	-
	60	.246	-.012	.147	.308	2.69	.228	.226	.124	.155	**
	120	.245	-.017	.103	.214	2.78	.228	.227	.086	.110	*
4	15	.190	-.001	.328	.355	2.72	.189	.183	.259	.418	**
	30	.198	-.009	.296	.138	2.69	.189	.186	.180	.296	**
	60	.173	.016	.180	.259	2.83	.189	.187	.126	.209	*
	120	.188	.001	.131	.210	3.42	.189	.188	.088	.148	**
5	15	.038	.049	.375	.182	2.66	-.013	-.013	.267	.335	**
	30	-.014	-.001	.249	.148	2.93	-.013	-.013	.186	.237	*
	60	-.025	-.012	.169	.293	2.65	-.013	-.013	.130	.167	**
	120	-.027	-.014	.118	.212	2.75	-.013	-.013	.092	.118	**
6	15	.040	.028	.386	.186	2.27	.068	.066	.266	.357	**
	30	.043	.025	.245	.192	2.66	.068	.067	.185	.252	**
	60	.029	.039	.186	.021	3.04	.068	.067	.130	.178	**
	120	.019	.049	.124	.334	3.33	.068	.068	.091	.126	**
7	15	.190	-.004	.313	.446	2.79	.186	.180	.260	.325	-
	30	.166	.020	.227	.294	2.64	.186	.183	.180	.230	**
	60	.190	.004	.165	.170	2.89	.186	.184	.126	.162	*
	120	.188	-.002	.111	.060	2.89	.186	.185	.088	.115	-
8	15	.180	-.038	.324	.064	2.59	.142	.137	.263	.319	*
	30	.177	-.035	.216	.162	3.05	.142	.140	.182	.225	*
	60	.170	-.028	.156	.075	2.53	.142	.141	.128	.159	**
	120	.171	-.029	.114	.104	2.96	.142	.142	.090	.113	**
9	15	.163	-.014	.398	.297	2.59	.149	.144	.262	.342	**
	30	.104	.045	.256	.167	2.66	.149	.147	.182	.242	**
	60	.130	.019	.174	.049	2.73	.149	.148	.127	.171	**
	120	.142	.007	.132	.084	2.64	.149	.149	.089	.121	*

eight of the nine differences were less than .01, the exception being item 4, with a difference of .017. Although the standard deviations yielded by samples of size 15 showed larger discrepancies from the large sample values (approximately .03) the pattern of discrepancies was similar to that obtained from an internal criterion.

General agreement between the mean value of biserial r based upon an external criterion and the mean product moment correlation coefficient existed, however, the relative magnitude of the two means varied with sample size and population value. The differences within pairs of mean values were all less than .05. The standard deviations of biserial r based upon an external criterion were consistently larger than those of the product moment correlation coefficient as was the case when the internal criterion was employed.

Kolmogorov-Smirnov goodness of fit tests were performed to test the discrepancy between the sampling distribution of biserial r obtained when an external criterion was employed and the sampling distribution of the product moment correlation coefficient. Of the 36 goodness of fit tests performed only 6 indicated no significant difference between the observed and the theoretical distribution. The non-significant results were distributed somewhat over the possible sample sizes, 1, 1, 1, and 2 non-significant differences were found for sample sizes 15, 30, 60 and 120 respectively. The goodness of fit tests indicated the observed distribution of biserial r did not uniformly follow that of the product moment correlation coefficient.

The sampling characteristics of biserial r obtained from internal and external criterion scores appeared to be similar. The bias with respect to both the population value and the mean product moment correlation coefficient was similar for the various sample sizes and laced a systematic pattern of differences. The standard deviations of biserial r were close to the large sample values and had a common pattern of relationship to the standard deviation of the product moment correlation coefficient for both of the criteria employed. In neither case was a relationship discovered which would easily differentiate the empirical sampling distribution due to the use of an internal and external criterion in the computation of biserial r.

Summary of sampling characteristics of biserial r

The characteristics of the empirical sampling distributions of biserial r indicated biserial r had desirable characteristics when samples of size 15 to 120 were employed. The biases yielded by samples of size 15 ranged from .003 to .101 and for samples of 120 from .002 to .003 indicating the mean biserial r, with the exception of one case, .101, approximates the population value of biserial r. The observed standard deviation of biserial r closely followed that given by the large sample theory formula for all sample sizes and was quite close for samples as small as size 15. The most desirable characteristic of the sampling distribution of biserial r was its ability to maintain discriminating power as the item functioned among high or low scoring individuals. When samples of size 15 were employed, only a slight decrease in discrimination was noted as the absolute value of X_{50} increased. When samples of 120 were employed, no decrease in the mean value of biserial r was noted as the absolute value of X_{50} increased. The sampling characteristics of biserial r obtained from an internal criterion and external criterion appeared to be quite similar. The biases were somewhat randomly distributed over all sample sizes and population values and the standard deviation of the observed biserial r tended to follow that given by the large sample theory formula when either criterion was employed. Neither criterion produced a relationship which would easily differentiate the empirical results due to a change in criterion. In general, one can conclude the sampling characteristics of biserial r, although not following those of the product moment correlation coefficient, were desirable when samples as small as 15 were employed.

Product moment correlation coefficient estimated by the U-L 27% method

The observed sampling statistics of r_f based upon an internal criterion are presented in table 7, accompanied by the bias, $\bar{r}$, σ_r, and the results of the Kolmogorov-Smirnov goodness of fit tests. When samples of size 15 were employed, all but one of the biases were less than .07, the exception being item 4 which had a bias of .146 for samples of this size. When samples of 120 were employed, all of the biases were less than .04, and six of the nine items had biases less than .02. A decrease in bias with an increase in sample size was found for eight of the nine items, the exception being item 3 whose bias increased from .02 to .04 as sample size increased from 15 to 120. The observed biases were all in the direction opposite from those expected from correlation theory. The theoretical value of the mean product moment correlation coefficient was smaller than the population value of r_f for all of the nine items employed, thus, it appears as though the mean value of r_f obtained from Fan's tables

Table 7. Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of r_f for samples of size 15, 30, 60, 120, internal criterion.

Item	Sample Size	Observed					Theoretical			Goodness of Fit
		r_f	Bias	S_{r_f}	$\sqrt{\beta_1}$	β_2		$\bar{r}$	σ_r	D_r
1	15	.541	-.041	.347	1.26	4.15	.500	.486	.210	**
	30	.517	-.017	.247	.46	2.95	.500	.493	.142	**
	60	.510	-.010	.169	.04	2.62	.500	.497	.099	**
	120	.515	-.015	.117	.03	2.91	.500	.499	.066	**
2	15	.396	-.050	.381	.64	3.07	.340	.329	.241	**
	30	.385	-.045	.248	.37	3.70	.340	.335	.166	**
	60	.402	-.062	.174	.22	3.31	.340	.337	.116	**
	120	.367	-.027	.116	.08	2.68	.340	.339	.080	**
3	15	.500	.020	.348	.50	2.37	.520	.506	.205	**
	30	.505	.015	.242	.27	2.42	.520	.513	.139	**
	60	.499	.021	.160	.56	3.34	.520	.517	.096	**
	120	.481	.039	.105	.25	2.84	.520	.519	.064	**
4	15	.484	.146	.375	1.12	3.39	.630	.616	.174	**
	30	.599	.031	.238	2.08	7.81	.630	.623	.116	**
	60	.638	-.008	.158	1.38	4.50	.630	.627	.080	**
	120	.613	.017	.129	.53	2.25	.630	.629	.049	**
5	15	.297	-.067	.421	.53	2.60	.230	.222	.255	**
	30	.243	-.013	.274	.33	2.89	.230	.226	.177	**
	60	.267	-.037	.213	.07	2.58	.230	.228	.124	**
	120	.255	-.025	.132	.15	3.07	.230	.229	.086	**
6	15	.610	.020	.271	1.83	7.34	.630	.616	.174	**
	30	.642	-.012	.176	1.30	4.61	.630	.623	.116	**
	60	.638	-.008	.125	1.03	4.00	.630	.627	.080	**
	120	.638	-.008	.090	.78	3.44	.630	.629	.049	**
7	15	.335	-.035	.432	.65	2.84	.300	.290	.247	**
	30	.227	.023	.263	.25	3.73	.300	.295	.170	**
	60	.322	-.022	.190	.21	2.82	.300	.298	.119	**
	120	.312	-.012	.123	.11	2.85	.300	.299	.083	**
8	15	.735	-.065	.266	1.53	5.06	.670	.656	.161	**
	30	.744	-.074	.174	.83	3.19	.670	.662	.107	**
	60	.707	-.037	.143	2.07	15.83	.670	.667	.073	**
	120	.686	.016	.088	.06	3.17	.670	.669	.042	**
9	15	.559	-.009	.307	1.15	4.36	.550	.536	.197	**
	30	.581	-.031	.224	.93	3.40	.550	.543	.133	**
	60	.560	-.010	.175	1.01	4.45	.550	.547	.092	**
	120	.550	.000	.115	.15	2.52	.550	.549	.060	**

slightly over estimates the mean product moment correlation coefficient.

The standard deviations of r_f obtained from the empirical data were consistently greater than the theoretical standard deviation of the product moment correlation coefficient; the differences ranges from .04 to .20 for all items and sample sizes. The observed standard deviations did, however, exhibit a marked decrease in magnitude from approximately .38 for samples of size 15 to .100 for samples of size 120. The values of r_f obtained from the empirical data also were considerably more variable than would be yielded by the sampling distribution of the product moment correlation coefficient. The discrepancies were large when samples of size 120 were used and extremely large when samples as small as size 15 were employed.

Although the observed values of the mean value of r_f and the standard deviation of r_f based upon the Fan tables deviate considerably from the theoretically expected values, these results should be anticipated somewhat; as the method is based upon the upper and lower 27% extreme groups. When samples of size 15 were used, only four subjects are available for the upper, and four subjects for the lower group. This means when 15 subjects were employed, there are only 25 possible combinations of proportions correct in the upper and lower 27% groups for use in finding the values of r_f. Even when samples of size 120 are involved, there is a relatively small finite number of combinations of Pu and Pl with which to locate values of r. Presumably when sample size increases, the number of subjects in the extreme groups becomes large enough to obtain resolution comparable with that of the correlation coefficient itself. In addition, there is considerable variability introduced by the sampling process itself; a sample of size n is drawn and from this sample two extreme groups are subsequently selected.

The empirical sampling distribution of r_f was tested against the sampling distribution of the product moment correlation coefficient by means of the Kolmogorov-Smirnov goodness of fit test. The results of the goodness of fit tests indicated all of the sampling distributions of r_f were significantly different from that of the product moment correlation coefficient. Thus, when samples of size 120 or less are employed in the extreme groups method of estimating the product moment correlation coefficient; the resulting indices are not distributed as the product moment correlation coefficient.

Columns 6 and 7 of table 7 contain the $\sqrt{\beta_1}$ which is a measure of the skewness of the distribution and β_2 which is a measure of the kurtosis of the distribution. Inspection of columns 6 and 7 indicates considerable variability in these two statistics, which are functions of the higher order moments, and a rather wide spread of combination of values is noted for all items and all sample sizes. Although Pearson [37] gives formulas and values for the value of $\sqrt{\beta_1}$ and β_2 obtained from the distribution of the product moment correlation coefficient, statistical techniques do not exist whereby one can test the significance of the difference between the obtained $\sqrt{\beta_1}$ or β_2 and those of the product moment correlation coefficient. In the present case, however, the obtained values were considerably discrepant from those of the product moment correlation coefficient. In addition, the observed second moment is consistently different from that of the product moment correlation coefficient. The empirical evidence indicates quite clearly that the product moment correlation coefficient as estimated by the upper-lower 27% method was not distributed as the product moment correlation when samples of size 120 or less were employed.

The three dimensional relationships among r_f and the item parameters X_{50} and β are depicted in figures 10 and 11 for samples of size 15 and 120 respectively. When samples of size 15 were employed the surface exhibited a marked curvature as the item parameters increased in value. The decrease in the value of r_f was indicated by the pronounced downward curvature in the upper line of figure 10. The curvature was also present in the central and lower lines of the surface, indicating the reduction in discrimination was present at all levels of β. When sample size was increased to 120, the amount of curvature of the surface, although still present, was much less pronounced than was the case for samples of size 15. When a high degree of discrimination was present, as indicated by the upper line in figure 11, the curvature was relatively small. An item discrimination index should be unaffected by the criterion score level at which the item functions, a characteristic not possessed by r_f when samples as small as 120 or less are employed.

Figures 12 and 13 were obtained by plotting the observed standard deviation of r_f as a function of the corresponding population values of X_{50} and β. The elevation of the surface in figure 12 indicates the standard deviation of r_f was relatively large when samples of size 15 were used. In addition, a rather peculiar relationship was found between

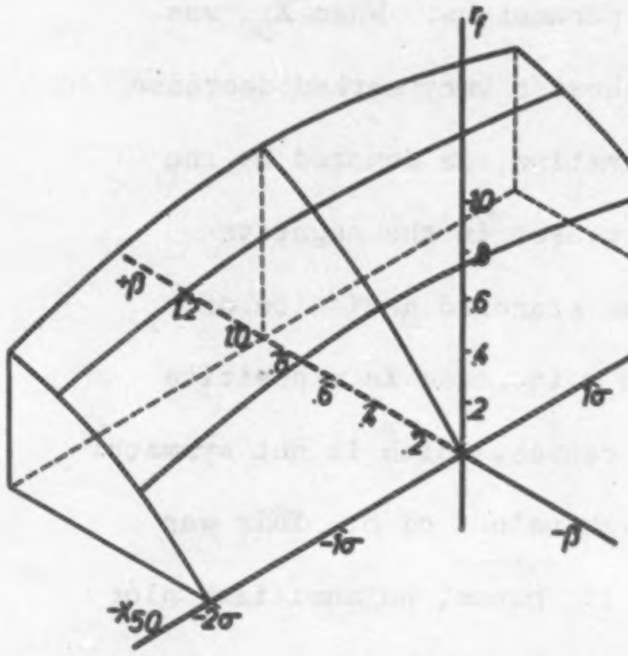


Figure 10. Observed relationship of r_I to the item parameters X_{50} and β as obtained from samples of size 15, internal criterion.

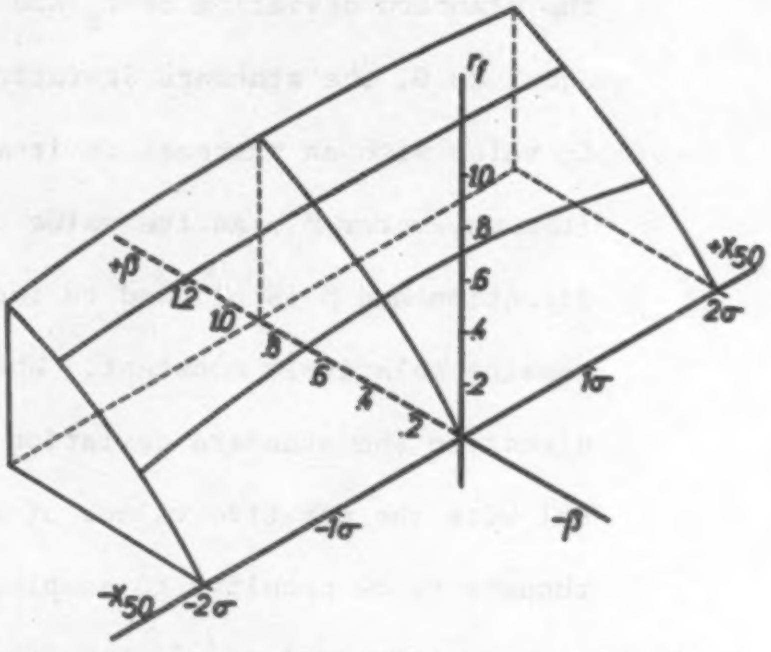


Figure 11. Observed relationship of r_I to the item parameters X_{50} and β as obtained from samples of size 120, internal criterion.

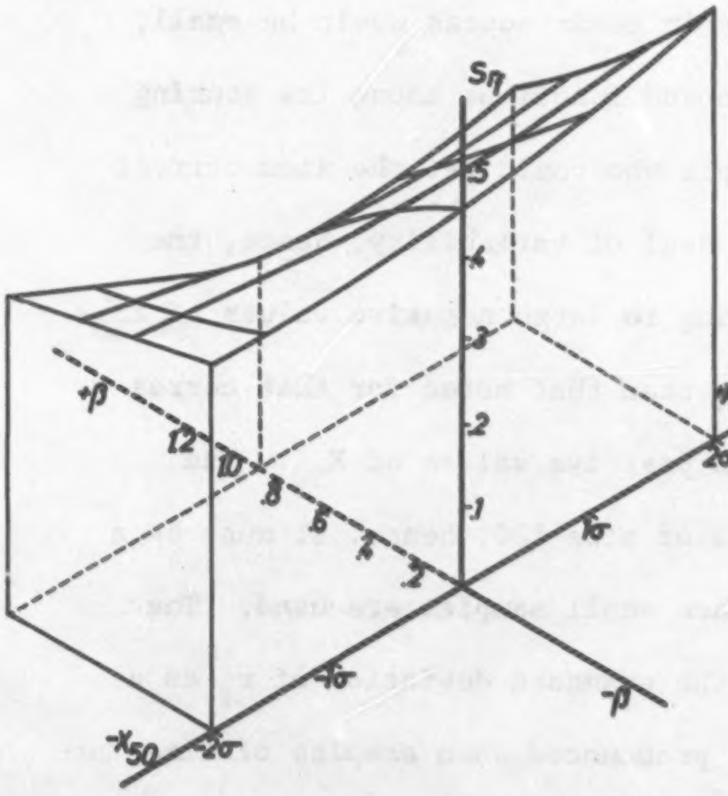


Figure 12. Observed relationship of the standard error of r_I to the item parameters X_{50} and β as obtained from samples of size 15, internal criterion.

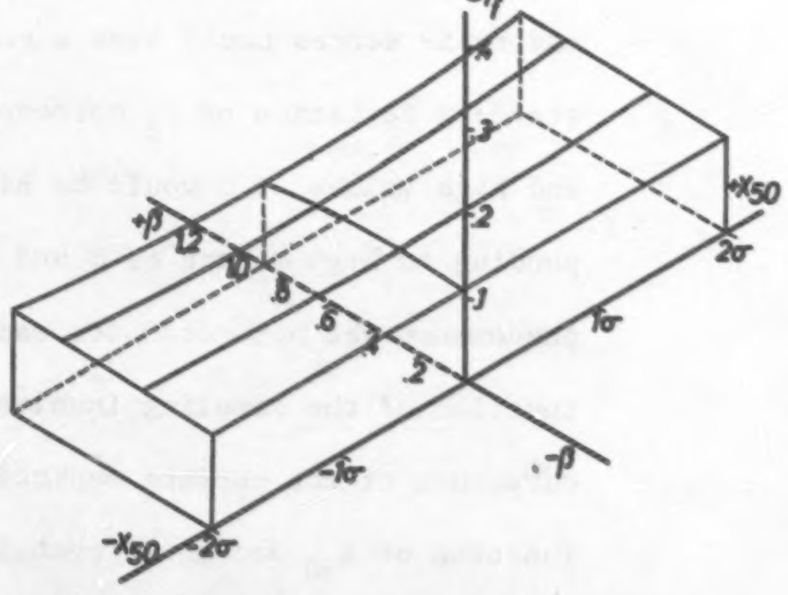


Figure 13. Observed relationship of the standard error of r_I to the item parameters X_{50} and β as obtained from samples of size 120, internal criterion.

the standard deviation of r_f and the item parameters. When X_{50} was equal to 0, the standard deviation of r_f shows a very marked decrease in value with an increase in item discrimination, as denoted by the item parameter β. As the value of X_{50} increases in the negative direction and β is allowed to increase, the standard deviation of r_f remains relatively constant. When X_{50} and β increase in a positive direction the standard deviation of r_f decreases, which is not symmetrical with the negative values of X_{50} and high values of β. This was thought to be peculiar to samples of size 15, hence, an auxiliary plot was made of samples; of size 30 and the same phenomenon was again noted. A possible explanation for this could be the following: when an item functions among high scoring individuals and has a high amount of discrimination, only a few high scoring people can possibly get the item correct, hence, the variability in their scores would be small; when an item has high discrimination and functions among low scoring people, there are many possible people who could get the item correct and their scores could have a great deal of variability, hence, the standard deviation of r_f corresponding to large negative values of X_{50} and high values of β would be higher than that noted for that corresponding to high values of β and high positive values of X_{50}. The phenomenon was not noted for samples of size 120; hence, it must be a function of the sampling involved when small samples are used. The curvature of the surface depicting the standard deviation of r_f as a function of X_{50} and β was much less pronounced when samples of size 120 were employed. When X_{50} was 0 and item discrimination allowed to increase, the standard deviation of r_f decreased slightly. As the absolute value of X_{50} was allowed to increase in conjunction with an increase in β, the standard deviation of r_f showed a slight increase in value. The resulting surface, however, was relatively uniform over the entire range of the item parameters.

The relationships among r_f, X_{50}, and β indicated the item discrimination index r_f has some undesirable sampling characteristics; namely, the index does not maintain its discriminating power when the item does not function near the mid-point of the criterion scale. The more discriminating the item was initially, the less discriminating it became as the item characteristic curve is moved away from the central portion of the criterion scale. This tendency, however, is decreased by an increase in sample size.

External Criterion

The population values of r_f based upon internal and external criterion were given in table 2. A large drop in the value of r_f occurred for all items when the criterion variable was changed from total test score to the grade point average. Those items having a high initial value of r_f exhibited the greatest loss in discriminating power. The sampling statistics of r_f based upon an external criterion are given in table 8 accompanied by the bias, $\bar{r}$, σ_r, and the results of the Kolmogorov-Smirnov goodness of fit tests. The biases of r_f were less than .062 for all items and sample sizes. The biases yielded by samples of size 15 ranged from .014 to .062, while those yielded by samples of 120 ranged from .000 to .045. Eight of the nine items exhibited a decrease in bias with increased sample size. Only the biases of item nine were greater when an external criterion was employed

Table 8: Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of r_f for samples of size 15, 30, 60, 120, external criterion.

Item	Sample Size	Observed					Theoretical			Goodness of Fit
		r_f	Bias	s_{r_f}	$\sqrt{\beta_1}$	β_2		$\bar{r}$	σ_r	D_r
1	15	.222	-.022	.480	.33	2.27	.200	.193	.258	**
	30	.235	-.035	.315	.25	3.17	.200	.197	.179	**
	60	.209	-.009	.193	.22	2.92	.200	.198	.125	**
	120	.219	-.019	.134	.06	3.79	.200	.199	.088	**
2	15	.216	-.046	.432	.14	2.43	.170	.164	.261	**
	30	.198	-.028	.288	.36	2.72	.170	.167	.181	**
	60	.160	.010	.177	.30	2.92	.170	.169	.127	*
	120	.170	.000	.120	.04	3.13	.170	.169	.089	-
3	15	.247	-.017	.434	.25	2.40	.230	.222	.255	**
	30	.235	-.005	.270	.34	3.22	.230	.226	.177	**
	60	.240	-.010	.176	.33	2.83	.230	.228	.124	**
	120	.242	-.012	.119	.07	3.54	.230	.229	.086	**
4	15	.135	-.045	.436	.23	2.34	.180	.174	.260	**
	30	.219	-.039	.410	.53	2.27	.180	.177	.180	**
	60	.173	.007	.280	.16	2.78	.180	.179	.126	**
	120	.135	.045	.170	.51	3.62	.180	.179	.088	**
5	15	.026	-.014	.464	.12	2.09	-.040	-.039	.267	**
	30	-.019	.021	.290	.02	2.78	-.040	-.039	.185	**
	60	-.046	-.006	.200	.08	3.00	-.040	-.040	.130	**
	120	-.060	-.020	.137	.13	2.57	-.040	-.040	.092	**
6	15	.023	-.023	.467	.14	1.84	.000	.000	.267	**
	30	.005	-.005	.325	.09	2.55	.000	.000	.186	**
	60	.006	-.006	.239	.20	3.63	.000	.000	.130	**
	120	-.017	..017	.140	.18	3.31	.000	.000	.092	**
7	15	.235	-.035	.463	.54	2.36	.200	.193	.258	**
	30	.192	.008	.294	.02	2.46	.200	.197	.179	**
	60	.188	.012	.191	.03	3.19	.200	.198	.125	**
	120	.189	.011	.132	.00	3.04	.200	.199	.088	**
8	15	.182	-.062	.388	.14	2.80	.120	.116	.264	**
	30	.173	-.053	.255	.13	3.48	.120	.118	.183	**
	60	.149	-.029	.187	.08	2.57	.120	.119	.128	**
	120	.157	-.037	.127	.13	2.60	.120	.120	.090	**
9	15	.147	-.037	.465	.27	2.17	.110	.106	.265	**
	30	.079	.031	.300	.06	2.56	.110	.108	.184	**
	60	.087	.023	.181	.04	2.49	.110	.109	.129	**
	120	.091	.019	.146	.08	2.85	.110	.110	.090	**

in the computation of r_f than those obtained from an internal criterion. Because of the smaller population values when the external criterion was employed, one would anticipate smaller biases in the observed mean of r_f.

The standard deviations of r_f obtained from the empirical data when an external criterion was employed were consistently greater than the theoretical standard deviations of the product moment correlation coefficient. The differences range from .031 to .230 for all items and sample sizes. The observed standard deviation did, however, exhibit a marked decrease in magnitude with increased sample size ranging from a value of approximately .46 for samples of size 15 to .14 for samples of size 120. The values of r_f obtained from the empirical data based upon the external criterion were considerably more variable than would be expected if r_f were distributed as the product moment correlation coefficient. The discrepancies were large when samples of size 120 were used and extremely large when samples of size 15 were employed. The relatively larger standard deviations obtained from an external criterion than those obtained from an internal criterion can be explained somewhat by figure 12. It was observed that as the population value of r_f decreased the standard deviation of r_f increased and as all population values based upon an external criterion are much less than those based upon an internal criterion, the former should have larger standard deviations.

General agreement between the mean value of r_f based upon an external criterion and the mean product moment correlation coefficient existed; however, the mean value of r_f tended to be larger than that of

2

the product moment correlation coefficient. The mean product moment correlation coefficient was exceeded by all sample sizes for seven of the nine items involved.

The Kolmogorov-Smirnov goodness of fit tests were performed to test the discrepancies between the sampling distribution of r_f obtained when an external criterion was employed and the corresponding sampling distribution of the product moment correlation coefficient. Of the 36 goodness of fit tests performed with samples of size 120, only item 2 showed no significant difference between the two distributions. One can conclude then, that the observed sampling distribution of r_f differed significantly from the theoretical distribution of the product moment correlation coefficient when an external criterion was employed.

The sampling characteristics of r_f obtained from an internal and external criterion score appear to be similar. The bias with respect to both the population value and the mean product moment correlation coefficient was similar for the various sample sizes. The standard deviations of r_f were consistently greater than those expected from the product moment correlation coefficient for both of the criteria. In neither case was a relationship discovered which would easily differentiate the empirical sampling distributions due to the use of an internal or an external criterion in the computation of r_f.

Summary of sampling characteristics of r_f

The sampling properties of r_f obtained from samples of 120 or less were not desirable. Only the mean value of the index closely approximated the theoretical values yielded by the sampling distribution of the product moment correlation coefficient. The observed standard deviations differed considerably from the theoretical for all items and sample sizes. The goodness of fit tests indicated 71 of the 72 empirical sampling distributions differed significantly from the corresponding distributions of the product moment correlation coefficient. The differences in sampling properties obtained through the use of an external and an internal criterion did not result in significant characteristics which would differentiate the sampling properties of r_f based upon either criterion. The results of the empirical sampling investigation indicate the upper-lower 27% method of estimating the product moment correlation coefficient does not have desirable sampling properties when samples of 120 or less are employed. The lower bound, for the use of this method, however, remains to be determined.

Cook's Index B

The observed sampling statistics of r_c based upon an internal criterion are presented in table 9, accompanied by the bias, $\bar{r}$, σ_r, and the results of the Kolmogorov-Smirnov goodness of fit tests. The bias of the mean values, with three exceptions for samples of size 15, was less than .05, for all items and sample sizes. Six of the nine items showed a decrease in bias as sample size increased, the remaining three items showed a slight increase in bias as sample size increased.

The reduction in the observed standard deviation of r_c was similar to that of the theoretical values obtained from the corresponding product moment correlation coefficient. When samples of size 120

Table 9: Empirical and theorietical sampling statistics and Kolmogorov-Smirnov goodness of fit tests for r_c for samples of size 15, 30, 60, 120, internal criterion.

Item	Sample Size	Observed					Theoretical			Goodness of Fit
		r_c	Bias	S_{r_c}	$\sqrt{\beta_1}$	β_2		$\bar{r}$	σ_r	D_r
1	15	.253	.006	.294	.054	2.60	.259	.250	.252	**
	30	.247	.012	.189	.180	3.10	.259	.255	.174	-
	60	.255	.004	.126	.044	2.65	.259	.257	.122	-
	120	.267	-.008	.090	.182	2.71	.259	.258	.085	-
2	15	.219	-.017	.277	.291	2.81	.202	.195	.258	*
	30	.229	-.027	.163	.034	3.36	.202	.199	.179	**
	60	.247	-.045	.132	.151	3.00	.202	.200	.125	**
	120	.226	-.024	.080	.021	2.77	.202	.201	.088	*
3	15	.308	.039	.258	.147	3.06	.347	.336	.240	**
	30	.319	.028	.182	.018	3.19	.347	.341	.165	**
	60	.312	.035	.128	.416	3.94	.347	.344	.115	**
	120	.324	.023	.086	.076	2.99	.347	.346	.080	**
4	15	.366	.097	.430	.352	1.60	.463	.450	.218	**
	30	.412	.051	.300	.058	2.45	.463	.457	.149	**
	60	.454	.009	.231	1.251	10.43	.463	.460	.103	**
	120	.448	.015	.149	.021	3.05	.463	.462	.070	**
5	15	.146	-.027	.293	.078	3.39	.119	.114	.264	**
	30	.109	.010	.192	.113	2.31	.119	.117	.183	-
	60	.116	.003	.144	.009	2.82	.119	.118	.128	-
	120	.123	-.004	.093	.224	2.93	.119	.118	.090	-
6	15	.279	.084	.356	.227	3.48	.363	.352	.237	**
	30	.313	.050	.221	.202	2.92	.363	.357	.163	**
	60	.305	.058	.166	.658	4.16	.363	.360	.114	**
	120	.321	.042	.103	.252	3.51	.363	.362	.078	**
7	15	.201	.007	.293	.324	2.64	.208	.201	.258	**
	30	.168	.040	.203	.243	2.81	.208	.205	.178	**
	60	.212	-.004	.140	.136	3.37	.208	.206	.125	-
	120	.204	.004	.091	.102	2.42	.208	.207	.087	-
8	15	.470	.062	.287	.174	2.56	.532	.518	.202	**
	30	.475	.057	.167	.420	3.53	.532	.525	.137	**
	60	.500	.032	.134	2.097	15.99	.532	.529	.094	**
	120	.516	.016	.082	.042	2.68	.532	.531	.062	**
9	15	.327	.020	.310	.158	3.23	.347	.336	.240	*
	30	.322	.025	.227	.053	2.66	.347	.342	.165	*
	60	.314	.033	.169	1.629	13.58	.347	.345	.115	**
	120	.316	.031	.095	.018	2.64	.347	.346	.079	**

were used, six of the nine observed standard deviations were within .01 of the theoretical product moment correlation coefficient; the differences found for the other three items were approximately .02 in magnitude. The greatest differences occurred with samples of size 15 and decreased as sample size increased. When samples of size 15 were used, five of the nine items were within .05 of the theoretical value, six of nine, eight of nine, and eight of nine were within .05 when samples of size 30, 60, 120, respectively were used.

General agreement between the mean value of r_c and the mean product correlation coefficient existed but was not consistent in detail; varying with sample size and/or population value of r_c. The differences within corresponding pairs of mean values, however, were less than .05 for all items except four and six, where rather large differences, .09, were noted for samples of size 15. The relative magnitude of the mean value of r_c and the mean of the product moment correlation coefficient varied from item to item. However, as sample size increased the mean value of r_c tended to approach that of the product moment correlation coefficient, although discrepancies existed for samples of size 120.

The Kolmogorov-Smirnov goodness of fit tests were performed to test the discrepancy between the sampling distribution of r_c and that of the product moment correlation coefficient. Eight of the thirty six tests of significance indicated the observed distribution did not differ significantly from the theoretical distribution of the product moment correlation coefficient. In the three items which conformed, the population values of r_c were relatively low, less than .25, however, others having equally as small population values did differ significantly from the distribution of the product moment correlation coefficient. No discernible relationship seems to exist which relates goodness of fit to the standard deviation of the distribution or to population value and/or sample size.

The three dimensional relationships among r_c and the item parameters X_{50} and β are depicted in figures 14 and 15 for samples of size 15 and 120 respectively. Inspection of figure 14 for samples of size 15 shows r_c has a tendency to decrease in discrimination with increased absolute values of X_{50}. There is a downward curvature of the surface along all lines of the surface indicating r_c loses discrimination as the item characteristic curve is shifted away from the mean of the criterion scale. As sample size increases, the tendency for r_c to drop in magnitude as the absolute value of X_{50} increases is reduced. In figure 15 the upper line of the surface shows a slight curvature for samples of size 120 indicating r_c was affected by shifts in value of the item parameter X_{50}.

The elevation of the surface in figure 16 indicates the standard deviation of r_c was relatively large when samples of size 15 were used. Along the line representing the value of X_{50} equal to 0, the standard deviation of r_c increases as discimination, β, increased. The increase was much more pronounced as the absolute value of X_{50} increased. At a value of X_{50} equal to $\pm 2\sigma$ the standard deviation of r_c increased very rapidly with increased values of β. The increase is indicated in figure 16 by the marked upward sweep of the surface at all levels of discrimination.

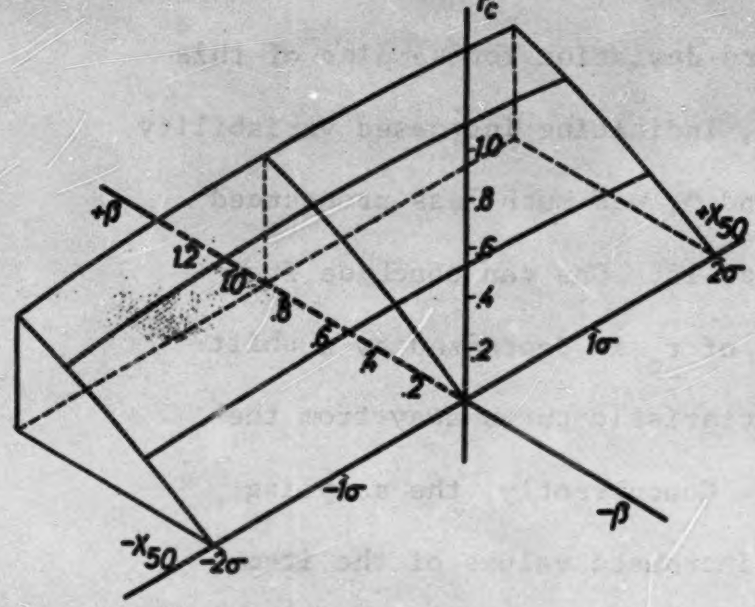


Figure 14. Observed relationship of r_c to the item parameters X_{50} and β as obtained from samples of size 15, internal criterion.

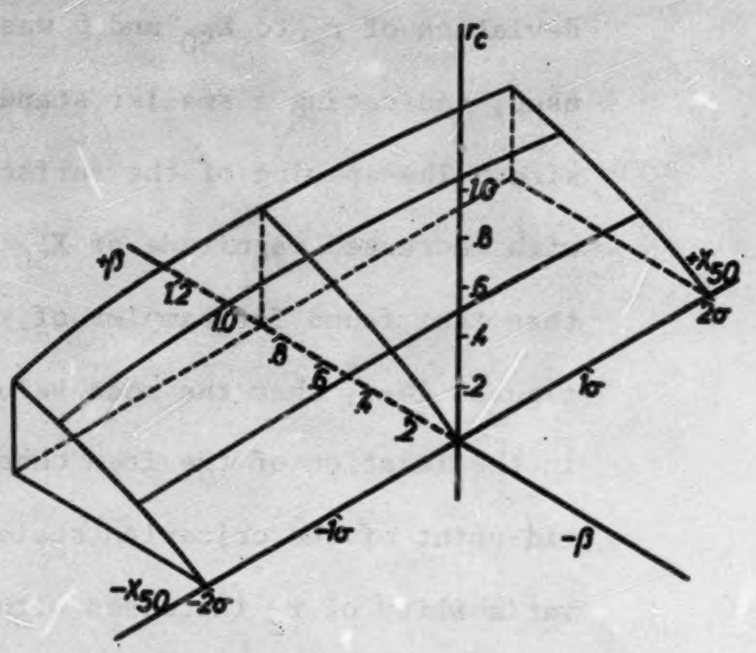


Figure 15. Observed relationship of r_c to the item parameters X_{50} and β as obtained from samples of size 120, internal criterion.

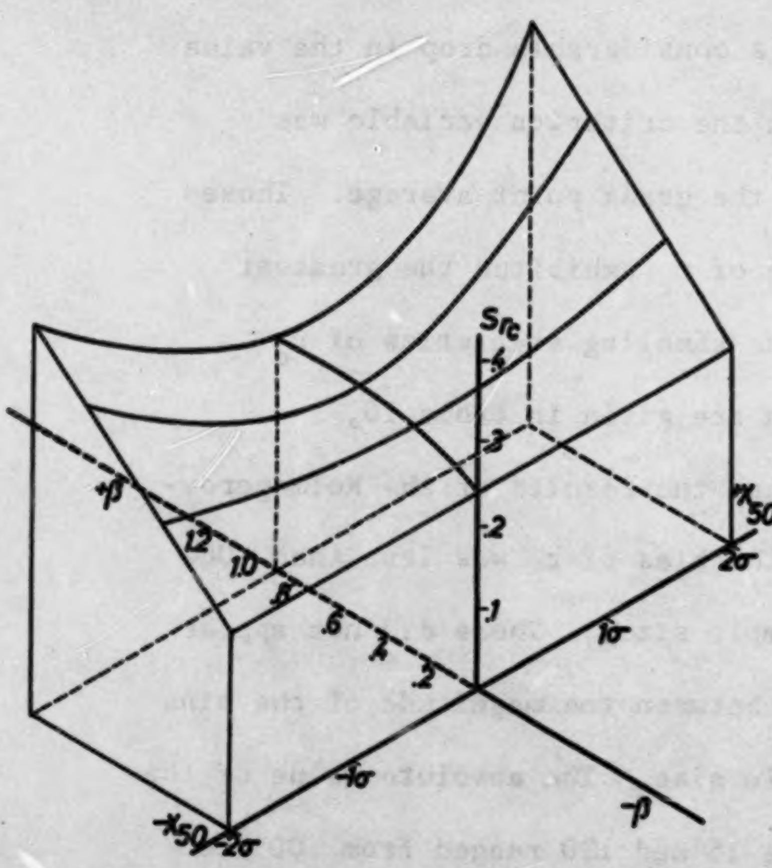


Figure 16. Observed relationship of the standard error of r_c to the item parameters X_{50} and β as obtained from samples of size 15, internal criterion.

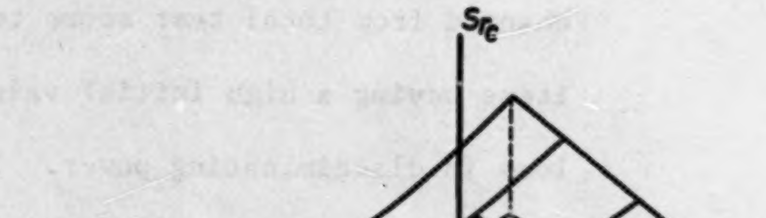


Figure 17. Observed relationship of the standard error of r_c to the item parameters X_{50} and β as obtained from samples of size 120, internal criterion.

The elevation of the surface, figure 17, relating the standard deviation of r_c to X_{50} and β was much less when samples of 120 were used, indicating a smaller standard deviation for samples of this size. The upswing of the surface, indicating increased variability with increased magnitude of X_{50} and β, was much less pronounced than that found for samples of size 15. One can conclude from figures 14-17 that the mean value of r_c is depressed by a shift in the location of the item characteristic curve away from the mid-point of the criterion scale. Concurrently, the sampling variability of r_c increases with increased values of the item parameters and/or a reduction in sample size.

External criterion

The population values of r_c based upon internal and external criteria were given in table 2. A considerable drop in the value of r_c occurred for all items when the criterion variable was changed from total test score to the grade point average. Those items having a high initial value of r_c exhibited the greatest loss in discriminating power. The sampling statistics of r_c based upon the external criterion are given in table 10, accompanied by the bias, $\bar{r}$, σ_r, and the results of the Kolmogorov-Smirnov goodness of fit tests. The bias of r_c was less than .064 for all population values and sample sizes. There did not appear to be a discernible relationship between the magnitude of the bias and the population value or sample size. The absolute value of the biases yielded by samples of size 15 and 120 ranged from .003 to .054 and from .004 to .062 respectively. Four of the nine items

Table 10: Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of r_c for samples of size 15, 30, 60, 120, external criterion.

Item	Sample Size	Observed					Theoretical			Goodness of Fit
		r_c	Bias	S_{r_c}	$\sqrt{\beta_1}$	β_2		$\bar{r}$	σ_r	D_r
1	15	.099	.003	.302	.337	3.08	.102	.098	.265	-
	30	.114	-.012	.206	.045	3.13	.102	.100	.184	-
	60	.111	-.009	.134	[illegible]	2.86	.102	.101	.129	-
	120	.114	-.012	.094	[illegible]	2.74	.102	.102	.091	-
2	15	.100	.006	.293	.102	2.78	.106	.102	.265	**
	30	.076	.030	.194	.242	2.55	.106	.104	.184	**
	60	.093	.013	.122	.153	3.26	.106	.105	.129	-
	120	.098	.008	.089	.002	3.14	.106	.106	.090	-
3	15	.116	.054	.267	.161	2.77	.170	.164	.261	**
	30	.135	.035	.176	.061	2.82	.170	.167	.181	**
	60	.151	.019	.122	.109	2.76	.170	.169	.126	**
	120	.166	.004	.083	.030	3.27	.170	.169	.089	-
4	15	.029	.005	.256	.179	6.56	.034	.033	.267	**
	30	.063	-.029	.230	1.513	6.26	.034	.033	.186	**
	60	.076	-.042	.147	.449	2.61	.034	.034	.130	**
	120	.048	-.014	.113	.326	2.36	.034	.034	.092	*
5	15	-.024	-.026	.302	.348	3.13	-.050	-.048	.267	*
	30	-.049	-.001	.197	.511	4.37	-.050	-.049	.185	-
	60	-.060	.010	.117	.142	2.57	-.050	-.050	.130	-
	120	-.060	.010	.091	.095	2.44	-.050	-.050	.091	-
	15	-.007	.051	.324	.327	4.38	.044	.042	.267	**
	30	-.004	.048	.221	.458	4.47	.044	.043	.185	**
	60	-.020	.064	.144	.102	2.51	.044	.044	.130	**
	120	-.018	.062	.091	.276	2.78	.044	.044	.091	**
7	15	.117	-.046	.298	.726	6.10	.071	.068	.266	**
	30	.094	-.023	.193	.208	3.40	.071	.070	.185	-
	60	.098	-.027	.139	.004	2.71	.071	.070	.130	-
	120	.086	-.015	.089	.027	3.17	.071	.071	.091	-
8	15	.117	-.026	.271	.136	2.90	.091	.088	.265	**
	30	.126	-.035	.189	.397	4.48	.091	.089	.184	**
	60	.108	-.017	.120	.105	2.64	.091	.090	.129	-
	120	.133	-.042	.094	.040	2.41	.091	.091	.091	**
9	15	.054	.046	.312	.023	3.64	.100	.096	.265	**
	30	.044	.056	.202	.180	2.96	.100	.098	.184	**
	60	.085	.015	.126	.171	2.74	.100	.099	.129	*
	120	.094	.006	.106	.067	2.86	.100	.100	.091	-

exhibited a consistent decrease in bias with increased sample size. The biases of five items increased as sample size increased, however, the increases were not particularly large.

The biases obtained from r_c using an internal and external criterion showed very similar behavior, both sets fluctuated inconsistently with sample size and population value of r_c. Four of the nine items based upon an external criterion had mean values of r_c greater than the population value, while only one item based upon the internal criterion had mean values of r_c greater than the population value. The biases based upon the external criterion were neither consistently larger nor consistently smaller than those obtained from the internal criterion. It should be noted, however, that the population values were always smaller when an external criterion was employed.

The observed standard deviation of r_c based upon an external criterion exhibited a reduction in magnitude with increased sample size similar to that of the product moment correlation coefficient. The differences between the observed and theoretical standard deviation were less than .021 when samples of 120 were used; seven of the nine differences were less than .006, the exceptions being items four and nine with differences of .021 and .015 respectively. Although the standard deviations based upon the external criterion yielded by samples of size 15 showed larger discrepancies from the theoretical values (approximately .03) than those yielded by an internal criterion the pattern of discrepancy was similar to that obtained from the internal criterion. It should be noted, however, that the standard deviations of r_c based upon an external criterion and samples of 120 were consistently closer to the theoretical values of r_c than were those based upon the internal criterion.

The computation of r_c using the external criterion produced values of the observed standard deviation which followed those of the product moment correlation coefficient and exhibited a pattern similar to that obtained when the internal criterion was employed. In both cases the magnitude of the standard deviation and the discrepancy between the observed and theoretical values decreased as sample size increased.

General agreement between the mean value of r_c based upon the external criterion and the mean product moment correlation coefficient existed, however, the relative magnitude of the two means varied with sample size and/or population value. Five of the nine items had mean values of r_c which were greater than the mean value of the product moment correlation coefficient.

The Kolmogorov-Smirnov goodness of fit tests were performed to test the discrepancy between the sampling distribution of r_c based upon the external criterion and that of the product moment correlation coefficient. Fifteen of the 36 goodness of fit tests indicated no significant difference between the observed sampling distribution of r_c and that of the product moment correlation coefficient. The agreement of the observed and theoretical distributions of r_c occurred as follows: 6 of 9, 5 of 9, 4 of 9, and 1 of 9 for samples of size 120, 60, 30, and 15 respectively. Item one which had a population value of .102 yielded observed sampling distributions of r_c which followed that of the product moment correlation coefficient for all four sample sizes. Items five and seven followed the theoretical distribution for samples of size 30, 60, and 120. Although the external criterion yielded empirical sampling distributions which tended to follow the theoretical distribution more often than those yielded by an internal criterion, this behavior was similar to that of the smaller population values obtained with the former criterion.

Summary of sampling characteristics of r_c

The empirical sampling characteristics of r_c possessed both desirable and undesirable characteristics, which were somewhat a function of sample size. The sampling results due to the internal and external criterion were very similar for all of the statistics employed in the analysis of r_c. No characteristics were found which would easily discriminate the sampling properties based upon the internal or the external criterion. The biases of r_c were relatively small for all sample sizes and item parameter values employed. The standard deviations of the observed r_c were uniformly very close to those of the corresponding product moment correlation coefficient. The standard deviation of r_c showed a marked decline in value as sample size increased as does the standard deviation of the product moment correlation coefficient. When the internal criterion was employed, eight of the 36 goodness of fit tests indicated the observed sampling distribution and the theoretical sampling distribution of the product moment correlation coefficient did not differ. When the external criterion was employed, fifteen of the 36 goodness of fit tests produced no significant differences. It would appear that small population values of r_c yield sampling distributions which follow those of the product moment correlation coefficient, although this was not uniformly the case. Figures 14 through 17 indicated the sampling properties of r_c had some undesirable characteristics when small samples were employed. Samples of size 15 possessed a rather high standard deviation and a decrease in discrimination as X_{50} and β increased in magnitude. The sampling characteristics were more desirable when samples of size 120 were employed; however, the mean value of r_c continued to exhibit a slight decrease in discrimination as the absolute value of the item parameter X_{50} increased. The standard deviation of r_c based upon samples of size 120, increased somewhat as the absolute values of X_{50} and β increased; however, the increase was much less marked than that observed for samples of size 15. The results of the empirical sampling investigation suggest that the empirical sampling properties of r_c when based upon samples of size 120 for items having a small degree of discrimination are those of the corresponding product moment correlation coefficient. It should be noted, however, this index is the correlation of the observed item characteristic and the theoretically perfect item characteristic curves. It is not an item-criterion correlation as is the case with the other indices in the present investigation.

Delta P

The observed sampling statistics of delta P based upon an internal criterion are presented in table 11, accompanied by the bias, $\bar{r}$, σ_r, and the results of the Kolmogorov-Smirnov goodness of fit tests. The observed bias of the mean value of delta P decreased with increased sample size in six of the nine items, the remainder exhibited an increase in the bias with increased sample size. When samples of size 15 were used, all but one of the biases were less than .05, the exception being item 3 which had a bias of .09. When samples of 120 were employed

Table 11: Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of delta P for samples of size 15, 30, 60, 120, internal criterion.

Item	Sample Size	Observed					Theoretical			Goodness of Fit	
		$\bar{P}d$	Bias	S_{Pd}	$\sqrt{\beta_1}$	β_2	Pd	$\bar{r}$	σ_r	D_B	D_r
1	15	.395	.016	.274	.181	2.55	.411	.399	.229	**	**
	30	.402	.009	.202	.173	3.30	.411	.405	.157	**	**
	60	.421	-.010	.139	.208	2.59	.411	.408	.109	**	**
	120	.446	-.035	.106	.194	2.66	.411	.410	.074	**	**
2	15	.330	.012	.303	.209	2.73	.342	.331	.241	**	**
	30	.363	-.021	.234	.324	3.60	.342	.337	.166	**	**
	60	.388	-.046	.168	.244	3.26	.342	.339	.116	**	**
	120	.359	-.017	.115	.020	2.68	.342	.341	.080	**	**
3	15	.429	.091	.310	.111	2.51	.520	.506	.205	**	**
	30	.476	.044	.229	.201	2.48	.520	.513	.139	**	**
	60	.488	.032	.162	.416	3.22	.520	.516	.096	**	**
	120	.475	.045	.108	.157	2.88	.520	.519	.064	**	**
4	15	.279	.033	.249	.315	2.87	.312	.302	.245	**	**
	30	.304	.008	.169	.221	3.54	.312	.307	.169	**	**
	60	.331	-.019	.136	.279	2.77	.312	.309	.118	**	*
	120	.328	-.016	.091	.077	2.65	.312	.311	.082	**	*
5	15	.234	-.036	.340	.216	2.87	.198	.191	.259	**	**
	30	.208	-.010	.228	.267	2.75	.198	.195	.179	**	**
	60	.224	-.026	.179	.083	2.77	.198	.196	.125	**	**
	120	.219	-.021	.114	.121	3.38	.198	.197	.088	**	**
6	15	.439	.012	.271	.021	2.91	.451	.438	.221	**	**
	30	.478	-.027	.190	.008	2.69	.451	.445	.151	**	**
	60	.462	-.011	.137	.201	2.82	.451	.448	.105	**	**
	120	.457	-.006	.092	.157	2.71	.451	.450	.071	**	**
7	15	.258	.024	.327	.216	2.90	.282	.273	.249	**	**
	30	.248	.034	.231	.128	2.22	.282	.278	.172	**	**
	60	.296	-.014	.176	.114	2.67	.282	.280	.120	**	**
	120	.291	-.009	.119	.010	2.86	.282	.281	.084	**	*
8	15	.638	.045	.286	.524	2.86	.683	.669	.156	**	**
	30	.703	-.020	.176	.486	3.08	.683	.677	.104	**	**
	60	.694	-.011	.138	1.458	10.23	.683	.680	.071	**	**
	120	.690	-.007	.093	.190	2.80	.683	.682	.040	**	**
9	15	.436	.014	.284	.062	2.83	.450	.437	.221	**	**
	30	.456	-.006	.217	.040	2.66	.450	.444	.151	**	**
	60	.451	-.001	.164	.405	3.90	.450	.447	.105	**	**
	120	.436	.014	.101	.035	3.02	.450	.449	.071	**	**

all of the biases were less than .05 and six of the nine items had biases less than .02.

The mean value of delta P was an amazingly good estimate of the mean value of the product moment correlation coefficient. The majority of differences between the mean delta P and $\bar{r}$ were .05 or less, the greatest difference being only .08, thus, within the range of population values studied, the mean value of delta P based upon 200 samples is a good estimate of the mean product moment correlation coefficient for samples of size 15, 30, 60, and 120. This does not, however, imply any individual value of delta P would necessarily fall close to the theoretical value of the product moment correlation coefficient.

The standard deviations of delta P obtained from the empirical data were consistently greater than the theoretical standard deviations of the product moment correlation coefficient based upon the population value of delta P. The differences between the two measures ranged from .00 to .105 for all items and sample sizes. The observed standard deviations did, however, exhibit a marked decrease in magnitude with increased sample size, ranging from approximately .3 for samples of size 15 to approximately .10 for samples of size 120.

The empirical sampling distribution of delta P was tested against the sampling distribution of the product moment correlation coefficient and against that of the binomial distribution by means of the Kolmogorov-Smirnov goodness of fit test. The results of the goodness of fit test indicated all of the sampling distributions of delta P were significantly different from that of the product moment correlation coefficient and from that of the binomial distribution. Columns six and seven of table 11 contain the $\sqrt{\beta_1}$ and β_2 which are measures of the skewness and kurtosis of the distribution. Inspection of columns six and seven reveals considerable variability in these two functions of the high order moments, and a rather wide spread in combination of values.

The three dimensional relationships among delta P and the item parameters X_{50} and β are depicted in figures 18 and 19 for samples of size 15 and 120 respectively. When samples of size 15 were employed the surface exhibited a marked curvature as the item parameters increased in value. The decrease in the value of delta P is indicated by the pronounced downward curvature in the upper line of figure 18. The curvature is also present in the central and lower lines of the surface, indicating reduction in discrimination is present at all levels of β. When sample size was increased to 120, figure 19, the curvature of the surface was still present, although somewhat reduced from that observed for samples of size 15. An item discrimination index with desirable characteristics should be unaffected by the criterion score level at which the item functions; a characteristic not possessed by delta P when samples of 120 or less were employed.

Figures 20 and 21 were obtained by plotting the observed standard deviation of delta P as a function of the corresponding population values of the item parameters X_{50} and β. When samples of size 15 were employed, the standard deviation of delta P exhibited some rather peculiar characteristics. If X_{50} equals zero and β increases from a low to a high value, the standard deviation of delta P is increased.

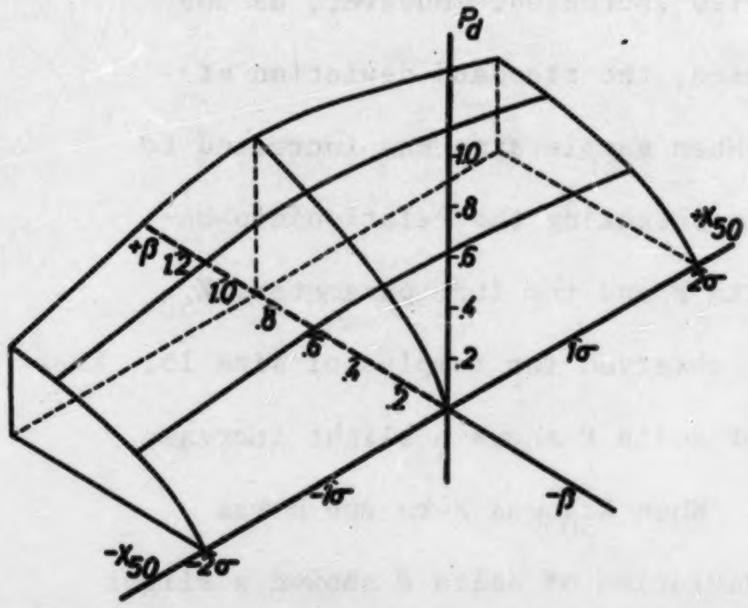


Figure 18. Observed relationship of delta P to the item parameters X_{50} and β as obtained from samples of size 15, internal criterion.

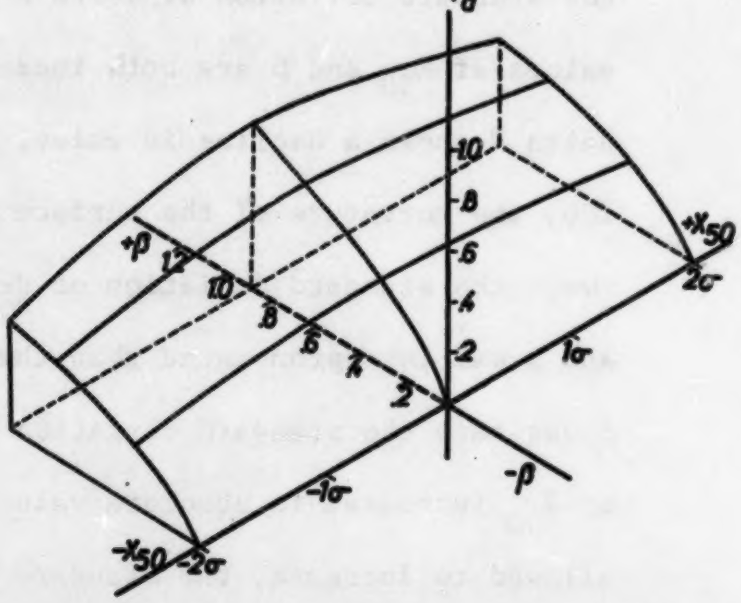


Figure 19. Observed relationship of delta P to the item parameters X_{50} and β as obtained from samples of size 120, internal criterion.

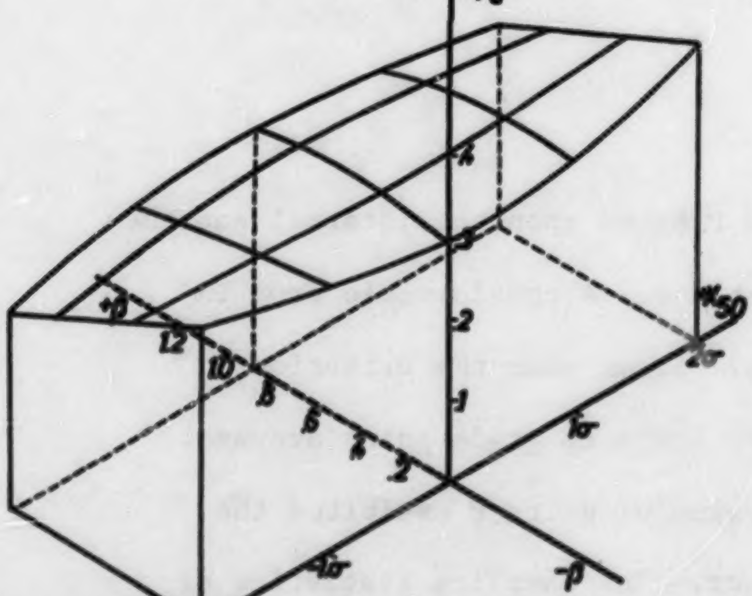


Figure 20. Observed relationship of the standard error of delta P to the item parameters X_{50} and β as obtained from samples of size 15, internal criterion.

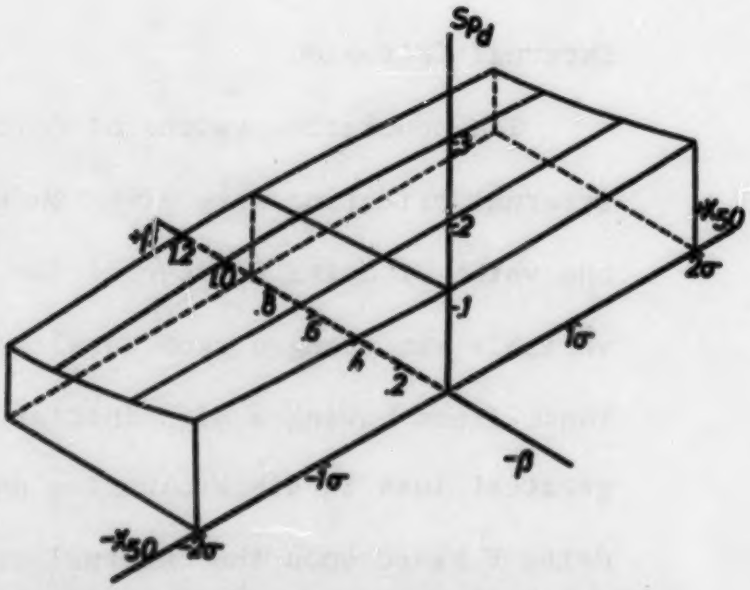


Figure 21. Observed relationship of the standard error of delta P to the item parameters X_{50} and β as obtained from samples of size 120, internal criterion.

When β has a value of zero and the absolute value of X_{50} is increased, the standard deviation of delta P also increases. However, as the values of X_{50} and β are both increased, the standard deviation of delta P shows a decline in value. When sample size was increased to 120, the curvature of the surface representing the relationship between the standard deviation of delta P and the item parameters X_{50} and β was less pronounced than that observed for samples of size 15. When β was zero the standard deviation of delta P shows a slight increase as X_{50} increased in absolute value. When X_{50} was zero and β was allowed to increase, the standard deviation of delta P showed a slight decrease in value which is the opposite of that noted for samples of size 15. With the exception of a slight increase in standard deviation as the amount of discrimination decreases, the surface formed by the standard deviation of delta P based upon samples of size 120 was relatively uniform, much more so than that observed for biserial r or r_c.

External Criterion

The population values of delta P based upon the internal and the external criterion were given in table 2. A considerable drop in the value of delta P occurred for all items when the criterion variable was changed from total test score to grade point average. Those items having a high initial value of delta P exhibited the greatest loss in discriminating power. The sampling statistics of delta P based upon the external criterion are given in table 12, accompanied by the bias, $\overline{r}$, σ_r, and the results of the Kolmogorov-Smirnov goodness of fit tests. The bias of delta P was less than

.042 for all population values and samples sizes. There did not appear to be a discernible relationship between the magnitude of the bias and the population value or sample size. The absolute value of the biases ranged from .00 to .042, .003 to .041, .002 to .023, and .007 to .033 for samples of size 15, 30, 60, and 120 respectively. Only two of the nine items exhibited a consistent decrease in bias with increased sample size. The remainder of the items either increased or fluctuated as sample size varied. The biases obtained from delta P using the internal and the external criterion showed similar general behavior, both sets fluctuated inconsistently with sample size and population value of delta P. It should be noted, however, the population values were always smaller when the external criterion had been employed.

The observed standard deviation of delta P based upon the external criterion exhibited a reduction in magnitude with increased sample size similar to that of the product moment correlation coefficient. However, the observed standard deviations of delta P were considerably larger for all sample sizes and items than the corresponding standard deviations of the product moment correlation coefficient. When samples of size 15 were used in conjunction with an external criterion, the standard deviation of delta P was approximately .32. When samples of size 120 were used, the observed standard deviation of delta P was approximately .11. The observed standard deviations of delta P based upon an external criterion tended to be larger than those obtained from an internal criterion, even though the population values of delta P were considerably smaller when the former was employed.

The Kolmogorov-Smirnov tests were performed to test the discrepancy between the sampling distributions based on the external criterion and the corresponding sampling distribution of the product moment correlation coefficient and the binomial distribution. When compared with the sampling distribution of the product moment correlation coefficient, all of the obtained empirical distributions differed significantly from that of the product moment correlation coefficient. All of the empirical distributions also differed significantly from the binomial distribution.

The sampling characteristics of delta P obtained from an internal and external criterion differ somewhat, but in a manner consistent with the observed difference in the population values. The bias of delta P obtained from an external criterion tended to be smaller than those obtained from an internal criterion, the standard deviation obtained from an external criterion was larger than those obtained from an internal criterion. Figures 20 and 21 in which the standard deviation of delta P is related to the item parameter β showed an increase in variability as discimination decreases. Thus, the results obtained from the two criteria are consistent with this relationship. Neither criterion scale yielded a relationship which would easily differentiate the sampling results due to the use of an internal or an external criterion in the computation of delta P.

Summary of sampling characteristics of delta P.

The analysis of the empirical sampling distributions of delta P indicated rather poor sampling characteristics for samples as large as 120. None of the empirical sampling distributions followed the

Table 12: Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of delta P for samples of size 15, 30, 60, 120, external criterion.

Item	Sample Size	Observed			Theoretical					Goodness of Fit	
		$\overline{Pd}$	Bias	S_{Pd}	$\sqrt{\beta_1}$	β_2	Pd	$\overline{r}$	σ_r	D_B	D_r
1	15	.169	.003	.336	.260	2.83	.172	.166	.261	**	**
	30	.182	-.010	.231	.234	3.14	.172	.169	.181	**	**
	60	.175	-.003	.165	.137	3.06	.172	.170	.127	**	*
	120	.186	-.014	.112	.155	2.79	.172	.171	.089	**	**
2	15	.178	-.006	.357	.084	2.62	.172	.166	.261	**	**
	30	.183	-.011	.266	.462	2.78	.172	.169	.181	**	**
	60	.154	.018	.170	.310	2.91	.172	.170	.126	**	*
	120	.165	.007	.117	.037	3.20	.172	.171	.089	**	*
3	15	.218	.009	.362	.011	2.42	.227	.219	.256	**	**
	30	.218	.009	.244	.286	3.02	.227	.223	.177	**	**
	60	.229	-.002	.168	.329	2.91	.227	.225	.124	**	**
	120	.236	-.009	.118	.093	3.50	.227	.226	.086	**	*
4	15	.064	.010	.206	.021	3.26	.074	.071	.266	**	**
	30	.092	-.018	.170	.046	2.62	.074	.073	.185	**	**
	60	.071	.003	.107	.058	2.76	.074	.073	.130	**	*
	120	.062	.012	.074	.167	2.98	.074	.074	.091	**	**
5	15	.022	-.042	.343	.092	2.73	-.020	-.019	.267	**	**
	30	-.024	.004	.237	.271	2.96	-.020	-.020	.186	**	**
	60	-.040	.020	.165	.142	2.63	-.020	-.020	.130	**	**
	120	-.051	.031	.116	.109	2.59	-.020	-.020	.092	**	**
6	15	.015	.000	.312	.191	2.48	.015	.014	.267	**	**
	30	.012	.003	.217	.025	2.82	.015	.014	.186	**	*
	60	.004	.011	.160	.193	3.23	.015	.015	.130	**	**
	120	-.013	.028	.100	.313	3.36	.015	.015	.092	**	**
7	15	.170	.003	.308	.273	2.46	.173	.167	.261	**	**
	30	.146	.027	.227	.279	2.44	.173	.170	.181	**	**
	60	.162	.011	.162	.106	2.92	.173	.172	.126	**	*
	120	.164	.009	.113	.019	3.07	.173	.172	.089	**	*
8	15	.159	-.039	.313	.043	2.94	.120	.116	.264	**	**
	30	.161	-.041	.228	.015	3.01	.120	.118	.183	**	**
	60	.143	-.023	.178	.065	2.58	.120	.119	.128	**	**
	120	.153	-.033	.123	.151	2.61	.120	.120	.090	**	**
9	15	.122	-.032	.350	.059	2.79	.090	.087	.265	**	**
	30	.064	.026	.229	.026	2.81	.090	.088	.184	**	**
	60	.072	.018	.148	.151	2.58	.090	.089	.129	**	**
	120	.077	.013	.121	.051	2.76	.090	.090	.091	**	**

sampling distributions of the product moment correlation coefficient or that of the binomial distribution.

Because delta P and r_f are both obtained from the extreme groups, they share a common limitations. Despite the very small number of subjects involved in the calculation of delta P when samples of 15 were used, the mean value of delta P was a reasonably good estimate of the population value of delta P and of the mean product moment correlation coefficient.

The close agreement of these measures held true for all sample sizes, population values of delta P involved, and criteria. The observed standard deviation of delta P was consistently larger than that of the corresponding product moment correlation coefficient for both an internal and external criterion. In both cases the observed standard deviation decreased as sample size increased. An undesirable characteristic of delta P was illustrated in figures 18 and 19 in which the curvature of the surface indicated delta P lost discrimination as the absolute value of the item parameter X_{50} increased in magnitude Increasing sample size from 15 to 120 did little to alleviate this situation as when samples of 120 were employed, a marked decrease in the discriminating power of the index also accompanied an increase in the magnitude of the item parameter X_{50}. When samples of 120 were employed, the standard deviation of delta P was uniformly small for all pairs of item parameters and exhibited a slight decline in variability with increased absolute values of X_{50} and β.

The results of the empirical investigation of the sampling distribution of delta P indicated that for samples of size 120 or less the index does not possess particularly desirable characteristics.

Item Parameter X_{50}

Maximum likelihood estimates of the item parameter X_{50} were not obtained when samples of size 15 were employed due to arithmetic difficulties encountered in the computer program. As was indicated in equation (8), X_{50} is a function of the reciprocal of β, hence, approaches infinity as β approaches zero. Lack of discrimination is a relatively common characteristic of items, especially when small samples are employed. The lack of discrimination, however, results in unrealistically large estimates of X_{50} and the existence of an isolated large value of X_{50} in an empirical sampling distribution results in misleading values of the mean and standard deviation, hence, the extremely deviant values of X_{50} were removed from some items when samples of 30 and 60 were employed. Deviant values did not occur when samples of 120 and an internal criterion were used. When an external criterion was employed, it was necessary to remove some deviant values from most items and sample sizes.

The observed sampling statistics of the maximum likelihood estimates of the item parameter X_{50} based upon an internal criterion are presented in table 13. The biases observed ranged from .070 to .533, -.013 to .288, and -.003 to .182 for samples of size 30, 60, and 120 respectively. Although the biases appear large relative to those yielded by the item discrimination indices, it should be noted the values of X_{50} are in criterion variable standard deviations units.

Table 13: Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of the maximum likelihood estimate of the item parameter X_{50} for samples of size 30, 60, 120, internal criterion.

Item	Sample Size	$\overline{X}_{50}$	Bias	X_{50}	Sx_{50}	Goodness of Fit D_n
1	30	-.814	-.254	-1.068	4.142	**
	60	-1.004	-.064	-1.068	.579	**
	120	-.927	-.141	-1.068	.568	*
2	30	-.243	.193	-.050	1.386	**
	60	-.152	.102	-.050	.525	**
	120	-.118	.068	-.050	.347	-
3	30	-.061	.072	.011	.814	*
	60	.101	-.090	.011	.497	**
	120	-.032	.043	.011	.274	-
4	30	-2.286	.533	-1.753	2.613	**
	60	-1.932	.179	-1.753	1.853	**
	120	-1.765	.012	-1.753	.431	-
5	30	1.633	.235	1.868	3.551	**
	60	1.984	-.116	1.868	2.488	**
	120	1.874	-.006	1.868	1.156	**
6	30	1.742	-.452	1.290	2.604	**
	60	1.248	.042	1.290	.458	-
	120	1.293	-.003	1.290	.322	-
7	30	-.695	-.509	-1.204	4.165	**
	60	-1.492	.288	-1.204	1.369	**
	120	-1.386	.182	-1.204	.867	*
8	30	-.206	.070	-.136	.390	-
	60	-.149	.013	-.136	.245	-
	120	-.164	.028	-.136	.174	-
9	30	1.690	-.521	1.169	2.000	**
	60	1.338	-.169	1.169	.820	-
	120	1.274	-.105	1.169	.486	-

The observed standard deviations of $\hat{X}_{50}$ exhibited a marked decrease in magnitude with increased sample size. When samples of size 30 were used, the standard error of $\hat{X}_{50}$ was approximately 2.0 to 4.0 which dropped to approximately .3 to 1.1 for samples of size 120. It was not possible to obtain theoretical values of $\sigma_{x_{50}}$ as this quantity has not been derived for the random sampling situation. The large sample formulas given in equations 9 and 10 apply only to a given arrangement of criterion score levels; in the present situation the criterion score level employed was also a random variable.

The distribution of X_{50} is known to be that of the normal function, therefore, the goodness of fit tests were performed with this function. The observed sampling distribution of $\hat{X}_{50}$ followed that of the normal for six of the nine items when samples of 120 were employed. One of the distributions of $\hat{X}_{50}$ yielded by samples of size 30 and only three of those obtained from samples of size 60 followed the normal distribution.

External Criterion

The obtained sampling statistics of the maximum likelihood estimates of the item parameter X_{50} based upon an external criterion are presented in table 14. Omitting item 5 which had biases exceeding 33 for all sample sizes the biases ranged from .065 to 10.11, .031 to 4.64, and -.093 to 9.57 for samples of size 30, 60, and 120 respectively. The biases obtained were considerable larger than those yielded by an internal criterion, reflecting the larger population values of X_{50} based upon small population values of β. The obtained standard errors were very large for all but a few items. The goodness of fit

Table 14: Empirical and theorietical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of the maximum likelihood estimate of the item parameter X_{50} of size 30, 60, 120, external criterion.

Item	Sample Size	$\bar{x}_{50}$	Bias	Sx_{50}	X_{50}	Goodness of Fit D_n
1	30	-1.017	-1.489	2.626	-2.506	-
	60	.682	3.188	23.505	-2.506	**
	120	-1.869	-.637	1.658	-2.506	*
2	30	.005	.082	1.737	-.077	-
	60	-.171	.094	1.669	-.077	-
	120	-.235	.158	1.338	-.077	*
3	30	-.033	.065	1.881	.032	**
	60	.001	.031	1.642	.032	*
	120	.125	-.093	1.234	.032	**
4	30	-1.873	-4.246	18.508	-6.119	**
	60	-4.297	1.822	7.624	-6.119	**
	120	-5.483	-.636	15.287	-6.119	**
5	30	.580	-35.135	18.254	-34.555	**
	60	1.428	-39.737	10.007	-34.555	*
	120	-.753	-33.802	14.712	-34.555	*
6	30	.937	10.113	13.665	11.050	**
	60	6.412	4.638	42.726	11.050	**
	120	1.477	9.573	16.099	11.050	**
7	30	-.755	-1.415	5.481	-2.170	**
	60	-1.553	.617	7.884	-2.170	**
	120	-2.272	.102	4.486	-2.170	**
8	30	.370	.973	9.622	-.603	**
	60	-.192	-.411	8.838	-.603	**
	120	-1.382	.779	5.071	-.603	**
9	30	-2.214	6.480	20.244	4.266	**
	60	1.383	2.883	22.615	4.266	**
	120	3.494	.772	7.545	4.266	**

tests showed 24 of the 27 empirical distributions differed significantly from the hypothesized normal distribution.

The sampling characteristics of the estimate of the item parameter X_{50} were very unstable when an external criterion was used. The unreliable sampling behavior would appear to be a function of the lack of discrimination as all items possessed small values of β, hence, the probability of large estimates of X_{50} was high. When an internal criterion was employed and a low population value of β was present (item 5) undesirable sampling characteristics similar to those yielded by the external criterion were obtained.

Summary of sampling characteristics of estimates of X_{50}

The empirical sampling investigation of the maximum likelihood estimate of the item parameter X_{50} indicated the obtained values of $\hat{X}_{50}$ were normally distributed when samples of 120 were used and the population value of the item parameter β was .3 or greater. Large differences in the sampling proerties of $\hat{X}_{50}$ were noted when the criterion variable was changed from an internal to external criterion. The differences noted are felt to be a function of the small magnitude of the population values of β obtained when the latter criterion variable was employed. The lack of formulas for the theoretical mean and variance of estimates of X_{50} for the conditions imposed in the present results is a serious deficiency. The empirical sampling distributions obtained in the present study indicate the normal function is appropriate under these conditions.

Item Parameter β

Maximum likelihood estimates of the item parameter β were not obtained for samples of size 15 due to the arithmetic difficulties arising in fitting the normal ogive. The obtained sampling statistics of $\hat{\beta}$ based upon an internal criterion are given in table 15. The observed biases range from -.011 to -2.587, .022 to -.522, and .002 to .179 for samples of size 30, 60, and 120 respectively. Because the estimates of β have a possible range from positive to negative infinity, the magnitudes of the biases can be larger than those yielded by distributions of the correlation coefficient.

The observed standard deviation of $\hat{\beta}$ exhibited a marked decrease in magnitude as sample size increased. With the exception of item four, the standard deviations were from approximately .2 to 1.0 for samples of 120, from .3 to 1.0, from .2 to .5 for samples of 30 and 60 respectively. Item four yielded extremely biased and variable estimates of β which, hence, produced abberant values of the standard deviation of $\hat{\beta}$. The theoretical values of the standard deviations of β corresponding to the situation in the present study were unknown, hence, the empirical values could not be tested. The large sample formula given in equation (9) applies only to a given arrangement of criterion scores, thus, is not applicable.

The distributions of the obtained values of $\hat{\beta}$ were tested against the normal distribution. With the exception of item four, 22 of the 24 goodness of fit tests indicated no significant difference between the observed and theoretical distributions. All

Table 15: **Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of the maximum likelihood estimate of the item parameter β of size 30, 60, 120, internal criterion.**

Item	Sample Size	$\bar{\beta}$	Bias	S_β	β	Goodness of Fit D_n
1	30	.543	-.011	.290	.532	-
	60	.576	-.044	.209	.532	-
	120	.586	-.054	.208	.532	-
2	30	.463	-.105	.335	.358	-
	60	.454	-.096	.222	.358	-
	120	.402	-.044	.105	.358	-
3	30	.628	-.018	.337	.610	-
	60	.588	.022	.212	.610	-
	120	.577	.033	.143	.610	-
4	30	3.541	-2.587	12.900	.954	**
	60	1.476	-.522	1.842	.954	**
	120	1.133	-.179	.540	.954	**
5	30	.307	-.068	.302	.239	-
	60	.294	-.055	.182	.239	-
	120	.284	-.045	.121	.239	-
6	30	.931	-.217	.657	.714	**
	60	.770	-.056	.250	.714	-
	120	.731	-.017	.163	.714	-
7	30	.347	.013	.282	.360	-
	60	.356	.004	.190	.360	-
	120	.358	.002	.141	.360	-
8	30	1.193	-.264	.997	.929	*
	60	1.133	-.204	.465	.929	-
	120	.955	-.037	.204	.929	-
9	30	.798	-.162	.467	.636	-
	60	.710	-.074	.301	.636	-
	120	.663	-.027	.182	.636	-

three sample sizes yielded significant differences for item four. When samples of 60 and 120 were employed eight of the nine items followed the normal distribution. Thus, the normal distribution appears to be the appropriate model for the distribution of β based upon samples as small as 30 and an internal criterion. The only consistent exception was item four which had a high population value of β (.954) and a very low (-1.75) population value of X_{50}.

External Criterion

The obtained sampling statistics of the maximum likelihood estimates of the item parameter β based upon an external criterion are given in table 16. The biases were somewhat less variable; ranging from .002 to -.048, .004 to .037, and from .004 to .043 for samples of 30, 60, and 120 respectively; than those obtained from an internal criterion. The greater uniformity is due to the population values of β based upon an external criterion being less than .3 for all items.

The observed standard deviations of $\hat{\beta}$ based upon an external criterion exhibited a consistent decrease in magnitude as sample size increased. The observed standard deviation was approximately .3 when samples of 30 were used, .2 and .1 for samples of 60 and 120 respectively.

The goodness of fit tests for the empirical distributions of $\hat{\beta}$ based upon an external criterion indicated all of the obtained distributions followed the hypothesized normal distribution. This result is unique among all of the indices studied; no other index has had

Table 16: Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of the maximum likelihood estimate of the item parameter β of size 30, 60, 120, external criterion.

Item	Sample Size	$\bar{\beta}$	Bias	S_{β}	β	Goodness of Fit D_n
	30	.266	-.061	.309	.205	-
1	60	.214	-.009	.165	.205	-
	120	.237	-.032	.127	.205	-
	30	.203	.002	.278	.205	-
2	60	.187	.018	.161	.205	-
	120	.197	.008	.108	.205	-
	30	.243	-.009	.253	.234	-
3	60	.264	-.030	.189	.234	-
	120	.264	-.030	.112	.234	-
	30	.230	-.028	.443	.202	-
4	60	.227	-.025	.209	.202	-
	120	.173	.029	.143	.202	-
	30	-.017	.004	.317	-.013	-
5	60	-.009	-.004	.186	-.013	-
	120	-.009	-.004	.111	-.013	-
	30	.043	.024	.309	.067	-
6	60	.030	.037	.204	.067	-
	120	.024	.043	.111	.067	-
	30	.203	-.011	.263	.192	-
7	60	.210	-.018	.179	.192	-
	120	.198	-.006	.128	.192	-
	30	.192	-.048	.292	.144	-
8	60	.177	-.033	.172	.144	-
	120	.182	-.038	.121	.144	-
	30	.117	.032	.297	.149	-
9	60	.125	.024	.182	.149	-
	120	.143	.006	.131	.149	-

more than a small proportion of its empirical distributions follow the appropriate hypothesized distribution.

The empirical sampling distributions of $\hat{\beta}$ indicated small population values of β yielded similar biases and standard deviations when either an internal or external criterion was employed. The differences between the results based upon an internal and external criterion appear then to be due solely to the difference in magnitude of the population values of β.

Summary of sampling characteristics of estimates of β

Because the solution of the likelihood equations yields both X_{50} and β their properties should not be summarized separately. The most significant sampling property of the item parameters appears to be the degree to which the empirical distributions follow that of the normal distribution. The distribution of $\hat{\beta}$ followed the normal distribution for nearly all items, sample sizes and criteria. When samples of 120 were used and the population value of β was greater than .3 most of the distributions of $\hat{X}_{50}$ were normally distributed. The sampling properties of $\hat{X}_{50}$ and $\hat{\beta}$ are inversely related; when the magnitude of $\hat{\beta}$ decreases, its sampling properties improve, however, the sampling properties of $\hat{X}_{50}$ degrade.

Considering the unusual situation in which the values of $\hat{X}_{50}$ and $\hat{\beta}$ were computed: each raw score constituting an interval, use of a small random selection of intervals, and the observed proportion of correct response within an interval consisting mostly of 1's and 0's; the results obtained suggest considerable robustness for the maximum likelihood estimation of item parameters when samples as small as 30 are employed. Perhaps the grouping of criterion scores into broader intervals would improve the estimation and the empirical sampling characteristics of both $\hat{X}_{50}$ and $\hat{\beta}$.

Comparison of Sampling Properties of the Item Discrimination Indices

Analysis of the results of the empirical sampling investigation indicated both similarities and differences among the sampling distributions of the item discrimination indices. The two indices with the most desirable sampling characteristics were β and biserial r. Although computed under unfavorable conditions, the sampling properties of the maximum likelihood estimate of β proved to be the uniformly best of any index employed. The empirical distributions of β did not differ significantly from the normal distribution even when samples as small as 30 were employed. The biases of β were relatively small and the magnitude of the standard deviations of β decreased rapidly as sample size increased. The sampling properties of biserial r were the best of the correlation and heuristic indices employed. The biases were small for all items, sample sizes, and criteria employed. When samples of size 15 to 120 were employed, the empirical standard deviations of biserial r differed only slightly from the values yielded by the large sample formula. The agreement of the empirical results with the large sample formula results seems to indicate the large sample formula for the standard deviation of biserial r can be successfully employed for samples as small as size 15. An important characteristic of biserial r was its negligible loss in discriminating power as the item characteristic curve was shifted away from the mid-point of the criterion scale. When samples of 15 were employed, biserial r suffered only a slight loss in discriminating power. No loss in discriminating power existed under the same conditions when samples of 120 were employed. Among the correlational and heuristic indices, the ability to maintain discriminating power as the absolute value of X_{50} increased was unique to biserial r. When samples of size 15 were used, the remaining indices (r_f, r_c, delta P) exhibited a marked loss in discriminating power when the item characteristic curve was shifted away from the center of the criterion scale. Increased sample size reduced the loss for r_c and r_f but did not eliminate the behavior. The characteristics of biserial r: small bias, usefulness of the large sample formula for the standard error, and its ability to maintain discriminating power under diverse conditions suggests it is the best of the correlational or heuristic indices. It should be noted that there is a functional relationship between the value of biserial r and the item parameter β. Equation (12) below gives this relationship.

$$(12) \quad r_b = \frac{\beta}{\sqrt{1+\beta^2}}$$

Because of this relationship the desirable characteristics of biserial r occurring simultaneously with "good" sampling characteristics for β is not surprising.

Those item discrimination indices based upon the upper-lower 27% extreme groups yielded the least desirable sampling properties. The mean value of r_f tended to be greater than the population value, the opposite of that expected from correlation theory, and the obtained standard deviations of r_f were consistently larger than the corresponding theoretical standard deviations of the product moment correlation coefficient. The relationship of r_f to the item parameters X_{50} and β showed for small samples, r_f was unable to maintain discriminating power as the absolute value of X_{50} increased. Increasing sample size from 15 to 120 reduced the loss but did not completely eliminate it. The sampling characteristics of delta P followed those of r_f, exhibiting a larger variability than anticipated and a marked loss in discriminating power as the item characteristic curve was shifted away from the midpoint of the criterion scale. Increased sample size did not reduce the loss accompanying a shift in the item characteristic curve as had been the case with r_f. Delta P possessed two unique characteristics: one, the mean value of delta P was uniformly a good estimate of the theoretical mean value of the distribution of the corresponding product moment correlation coefficient, within the range of values employed (-.02 to .683); second, the variability of delta P decreased as the absolute values of both item parameters were increased. The sampling properties of indices based upon the extreme groups methods had characteristics: consistent bias, larger variability, and lack of ability to maintain discriminating power under diverse conditions; which would label these indices as the "worst" of those studied.

The sampling properties of Cook's index B were somewhat intermediate to those found for biserial r and r_f or delta P. The biases were small and the mean value of r_f was uniformly less than the population value, as would be indicated by correlation theory, and the observed standard deviations closely approximated those of the product moment correlation coefficient. The index, however, exhibited a loss in discriminating power as the item characteristic curve was shifted away from the center of the criterion scale. The loss was greater than that observed for biserial r but not as marked as that exhibited by r_f or delta P. Increasing sample size reduced the loss in discriminating power more effectively than was the case for delta P or r_f. When samples of 15 were employed, the standard deviation of r_c became extremely large when the absolute value of X_{50} and β were both increased. Using samples of size 120 reduced the observed variability considerably, however, the relationship to X_{50} and β was still present. The comparison of sampling properties of r_c with the remaining indices needs to be qualified by noting the index is not an item-criterion correlation.

The use of the sampling distribution of the product moment correlation coefficient as the model for the empirical distributions of the item discrimination indices was uniformly disappointing. Of the 288 Kolmogorov-Smirnov goodness of fit tests performed, 256 indicated a significant difference between the observed and theoretical distributions. The goodness of fit tests indicated quite clearly the item discrimination indices involved in the present study (r_b, r_f, r_c, delta P) could not be considered to be distributed as the correlation coefficient when samples of size 15, 30, 60, 120 are used. The empirical sampling statistics of the correlational and heuristic indices yielded many different combinations of bias, standard errors, $\sqrt{\beta_1}$ and β_2, but it is beyond the scope of the present study to determine the precise contributors to the discrepancy between the observed and theoretical sampling distributions.

THE HOYT RELIABILITY COEFFICIENT

Background to the Problem

The psychometric literature contains a large number of articles dealing with the derivation, computation, and the use of coefficients of internal consistency, i.e., the reliability coefficients. As was the case for item discrimination indices, a basis for interpreting the numerical values of these coefficients was not made explicit. Knowledge of the sampling properties of a reliability coefficient are necessary to make a meaningful interpretation of an obtained numerical value. The present study attempts to provide empirical evidence as to the sampling characteristics of one reliability coefficient for a single population value.

The method used in the present study to compute the reliability coefficient was the Hoyt Analysis of Variance method [21]. The method of computation is a two way analysis of variance in which one classification is the subjects and the other the items of the test.

The reliability coefficient is given by the following equation:

$$(13) \quad r_h = \frac{MS_{ind} - MS_{within}}{MS_{ind}}$$

In addition, Hoyt [21] showed the equivalence of his reliability coefficient and that obtained from the Kuder-Richardson formula 20 [30]. The sampling distribution of the coefficient, however, was not given.

In a paper which appeared slightly previous to Hoyt's, Jackson [23] presented the derivation of the "sensitivity" of a test. The sensitivity was the ratio of the standard deviation of the group to the standard error of measurement as given in equation 14 below.

$$(14)\quad r = \frac{\sigma_c}{\sigma}$$

Where σ_c is the standard deviation of distribution of ability

σ is the standard deviation of the distribution of the error of measurement.

He then showed the relationship between r and rho where rho is the population reliability coefficient.

$$(15)\quad \text{rho} = r^2 \,/\, (1 + r^2) \quad \text{or}$$

$$r = \sqrt{\text{rho} \,/\, (1-\text{rho})}$$

Substituting obtained values for the population values we have:

$$(16)\quad G = \sqrt{r/\ (1-r)}$$

Jackson's work [23] indicates the reliability coefficient is the square of a correlation coefficient; hence, its sampling distribution could be obtained from that of the product moment correlation coefficient.

Butler [3] considered the reliability coefficient to be the estimated proportion of the obtained variance which is true variance; thus, $1 - r_t$ represents the error or within individual variance. Since r_f and $1 - r_t$ were variances, their ratio should be F distributed.

$$(17)\quad F = r_t \,/\, (1-r_t)$$

If the sample size is N and there are n items, the degrees of freedom associated with r_t is N - 1 and the degrees of freedom associated with $1 - r_t$ is (N - 1) (n - 1). The ratio given by Butler [3] leads to peculiar results when one attempts to relate it to the sampling distribution of the reliability coefficient. The F ratio tests whether the reliability coefficient is significantly different from 0. For example in the present study, the population value of the Hoyt reliability coefficient was .906. Applying Butler's formula one obtains $F = .906/.094 = 9.63$ which corresponds to a cumulative area of the F distribution of approximately .9999 for 14 and 1106 degrees of freedom. Yet, the reliability coefficient involved was the true mean of the sampling distribution and should have an F ratio of approximately 1.0. A similar situation holds if one rewrites the Hoyt equation for the reliability coefficient as $F_{14,1106} = \frac{1}{1-r} = 1/(1-.906) = 10.63$, again a very high value of F for the degrees of freedom involved. The computer program for the F distribution (20) was employed to compute the cumulative area of the F distribution. However, the central F could not be manipulated to represent the empirical sampling distributions obtained in the present study.

The results given above indicated the central F distribution cannot be employed as a model for the sampling distribution of the Hoyt reliability coefficient when the population value of rho is not equal to 0. It appears possible, that the non-central F could serve as the model of the sampling distribution of the Hoyt reliability coefficient. Due to the lack of detailed tables or a computer program, the author was unable to pursue this avenue of approach.

The author was unable to locate a single published study of the sampling distribution of the reliability coefficient. The only information available was an unpublished study by Hoyt and McEachern [22] in which they computed a number of reliability coefficients based on samples of size 50 or less. Hoyt and McEachern found the mean value of the Hoyt reliability coefficient had a relatively small bias for samples of size 50 and the population values involved.

In the present study the Hoyt reliability coefficient was computed for each random sample as a part of the preliminary analysis in the item analysis program. The reliability coefficient (r_h) and the mean square within, were punched into Hollerith cards, however, only r_h was reported in the present study. When computed by the Hoyt method, the reliability coefficient involves both the number of items and the subjects in the item by individual matrix. Because the number of items in the instrument was fixed at 80 items, the degrees of freedom could be reduced only by reducing the sample size. Therefore, samples of size 15 were employed to investigate the effect of small degrees of freedom upon the sampling distribution of the reliability coefficient. The computer analysis produced 200 reliability coefficients for each of the four sample sizes.

Analysis of the empirical results

The observed sampling statistics of r_h; the mean, standard deviation, square roots of β, and β_2 are presented in table 17. The bias of the mean r_h decreased as sample size increased, falling from .012 for samples of size 15 to .003 for samples of size 60 and 120. The observed standard deviation of r_h also exhibited a decrease in magnitude; dropping from .04 for samples of size 15 to .012 for samples of size 120. When samples of size 120 were used the range of r_h was .87 to .93, indicating all of the obtained values were very close to the population value. Samples of size 60 possessed characteristics very similar to those of 120, having a bias, of .003, standard deviation of .018, and a range of .84 to .94.

Table 17: Empirical and theoretical sampling statistics and Kolmogorov-Smirnov goodness of fit tests of the Hoyt reliability coefficient for samples of size 15, 30, 60, 120.

Sample Size							Goodness of fit	
	$\bar{r}_h$	Bias	S_{r_h}	P_h	$\sqrt{B_1}$	B_2	D_r	$D\sqrt{r}$
15	.894	.012	.040	.906	1.324	5.941	**	**
30	.897	.009	.028	.906	1.207	4.868	**	**
60	.903	.003	.018	.906	.612	3.307	**	**
120	.903	.003	.012	.906	.240	2.688	**	**

The Kolmogorov-Smirnov goodness of fit tests were performed using the cumulative frequency distribution of the product moment correlation coefficient corresponding to the population value of r_h and that corresponding to the square root of the population value of r_h. In both cases the observed and theoretical distributions differed significantly for all sample sizes.

Figure 22 presents the empirical sampling distributions of r_h obtained from the four sample sizes. The dominate feature in the figure is the pronounced peaking of the curves and the decrease in negative skewness as sample size increased. The sharp reduction in the values of β_2 and the $\sqrt{\beta_1}$ in table 17 also indicated this phenomena. Unfortunately, Pearson's tables of the theoretical values of β_1 and β_2 [37] for the product moment correlation coefficient were not sufficiently detailed to directly compare the observed values.

Although the Kolmogorov-Smirnov goodness of fit tests indicated the observed frequency distribution of r_h did not follow that of the product moment correlation coefficient corresponding to the $\sqrt{r_h}$, the small bias and similar standard deviation of r_h prompted further inquiry. In figure 23 the cumulative frequency distributions of the observed r_h and the product moment correlation coefficient corresponding to r_h and the $\sqrt{r_h}$ were plotted for samples of size 120. The values of $\sqrt{r_h}$ were plotted in the following manner: If r_h equals .81 the cumulative area of the product moment correlation coefficient was obtained for r=.9, cumulative area was then plotted at .81 in figure 23.

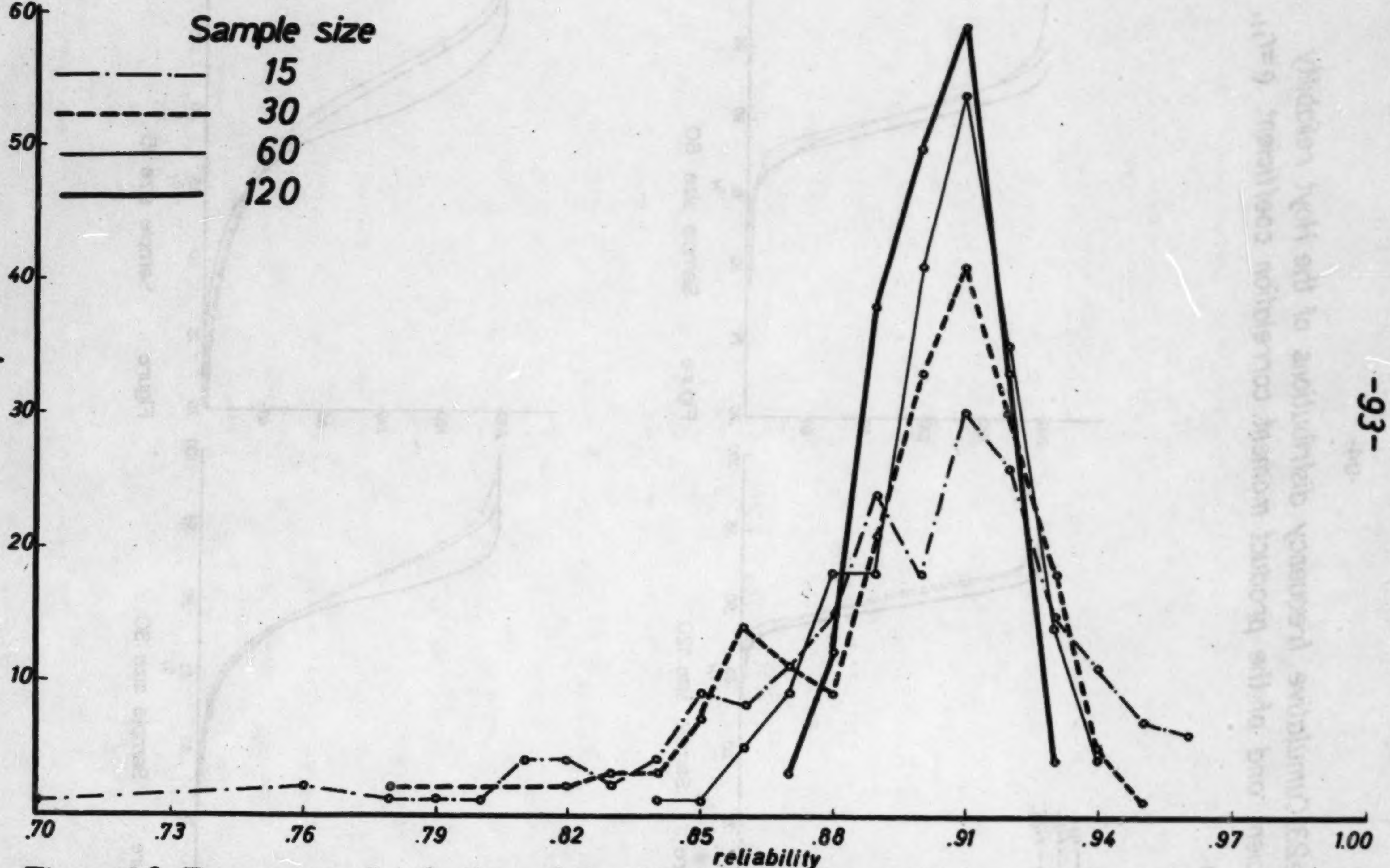


Figure 22. Frequency distribution of the Hoyt reliability coefficient obtained from samples of size 15, 30, 60, 120.

Figure 23. Cumulative frequency distributions of the Hoyt reliability coefficient and of the product moment correlation coefficient, $\rho = r_H$, $\rho = \sqrt{r_H}$.

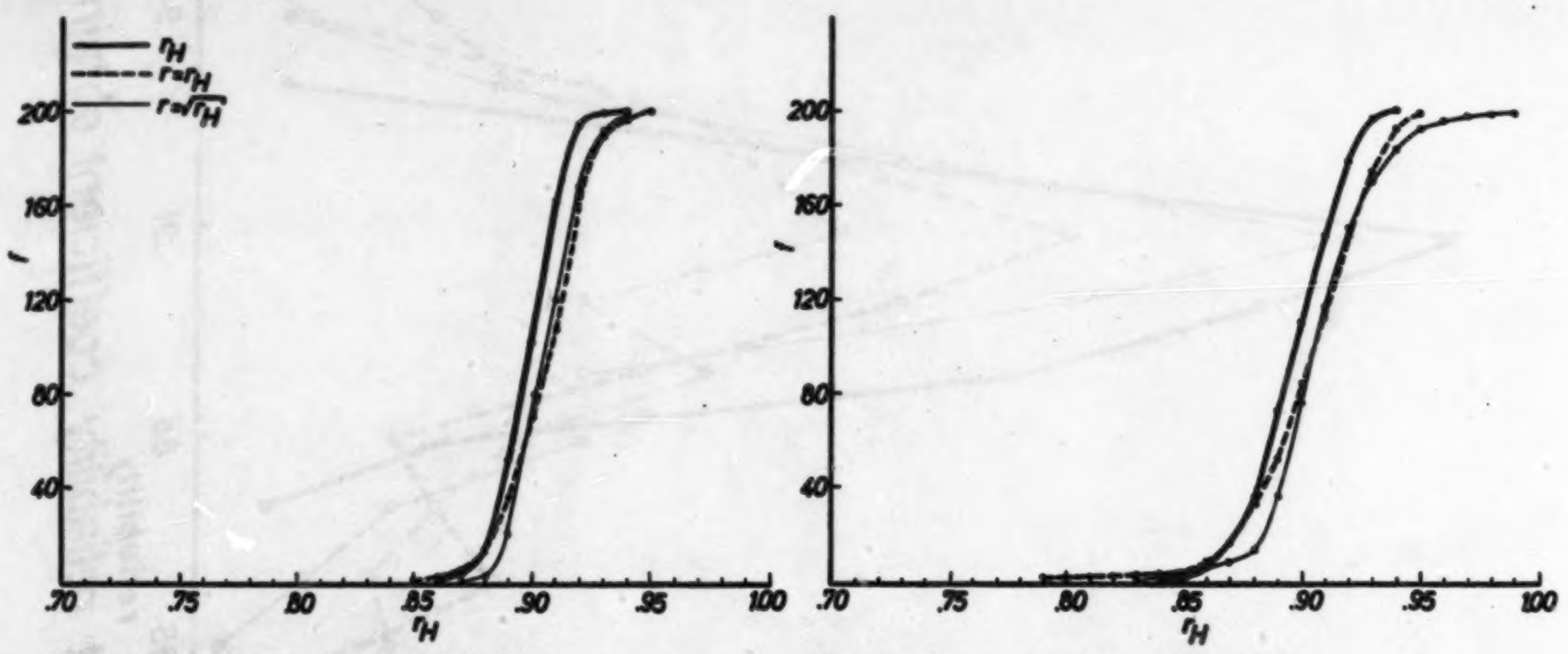


Figure Sample size 120.

Figure Sample size 60.

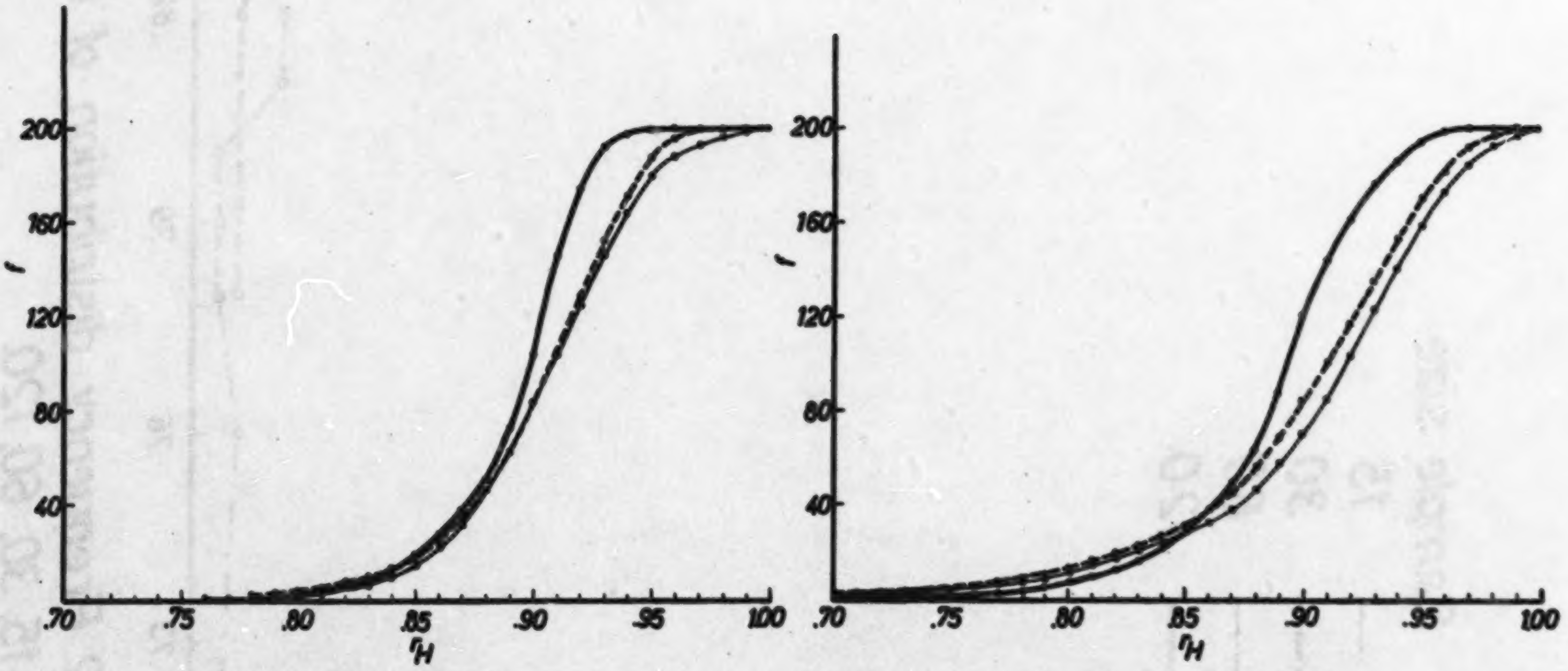


Figure Sample size 30.

Figure Sample size 15.

Except for a displacement toward higher values, the cumulative frequency distribution of the product moment correlation coefficient corresponding to the square root of r_h had a form nearly identical to that of the observed cumulative frequency distribution of r_h when samples of 60 and 120 were used. If one accounts for the slight bias, .003, in the samples of 120, the Kolmogorov-Smirnov tests indicated no significant difference between the cumulative frequency distribution of the product moment correlation coefficient corresponding to the $\sqrt{r_h}$ and the observed frequency distribution of the Hoyt reliability coefficient. Accounting for the bias in samples of 15, 30, and 60, however, did not result in agreement between the observed and theoretical distributions. The effect of this seemingly slight bias was considerable in the present situation, which can possibly be accounted for by the high population value of r_h and the restricted range of values in both the theoretical and observed distributions.

Summary of the Sampling Characteristics of the Hoyt Reliability Coefficient

The Hoyt reliability coefficient appears to have desirable sampling characteristics for the sample sizes, number of items, and the single population value employed in the present study. From a practical point of view, the bias of the mean value of r_h was small for samples as low as size 15 and negligible for samples of size 120. In addition, the range of obtained values was limited when samples of 60 and 120 were employed. In the present case, the results indicated that if the population value of the reliability coefficient were .906, a reasonably close estimate could be obtained from samples of size 60 or larger. When samples of 120 were employed, the distribution of the product moment correlation coefficient corresponding to the square root of the reliability coefficient and adjusted for a slight bias, did not differ significantly from the observed distribution of the Hoyt reliability coefficient.

The question which needs to be reconciled with the present data, however, is the effect of the number of items. In the present case the distribution of the product moment correlation coefficient based upon the sample size was employed; however, the Hoyt method is a function of both the number of items and individuals, thus leaving one in a quandry as to the effect of items. Perhaps the bias noted for samples of 120 is related to the effect of the number of items.

The F distribution did not prove to be a good model for the sampling distribution of the Hoyt reliability coefficient, however, a possibility exists that the non-central F is the appropriate model. On the basis of the present results, the distribution of the product moment correlation coefficient corresponding to the $\sqrt{r_h}$ appears to be a more likely model.

SUMMARY AND CONCLUSIONS

Meaningful interpretation of a statistic is dependent upon knowledge of its sampling distribution, the conditions which affect this distribution, and any pathological characteristics of the statistic. Although item discrimination indices have been used for many decades, knowledge of their sampling characteristics is limited. The present study was designed to ascertain the sampling characteristics of five item discrimination indices (r_b, r_f, r_c, delta P, and β), two associate statistics (difficulty and X_{50}), and a reliability coefficient, under the following conditions: two criteria (internal and external), four samples sizes (15, 30, 60, and 120), and nine items having selected item characteristic curves. The conditions established were felt to be sufficiently systematic to adequately investigate the sampling characteristics of the selected item discrimination indices.

The population data consisted of the responses to the items of the ACE-Q test, the test scores, and the freshman grade point averages of 747 university freshmen. Both the total test scores and the freshman grade point averages were normally distributed with a common mean and variance. Univac 1103 digital computer programs were used to item analyze random samples of four specified sizes drawn at random from the above population. The results of the 1103 analyses were then subjected to further data reduction and analysis by means of computer programs written for a Control Data 1604 computer.

The study was designed with five questions serving as the basic guide lines; the paragraphs below relate the results of the study to the original five questions.

1. What are the effects of sample size upon the sampling characteristics of the indices?

The only sampling characteristic which was consistently related to sample size was the standard error of the index. In all cases, the observed standard error decreased as the sample size increased; the magnitude of the decrease was a function of sample size and the population value of the particular index. The bias of the mean item discrimination index decreased with sample size for certain items, but not consistently enough to make it a clear-cut pattern as was the case for the standard deviation.

Sample size and the degree to which the distribution of the index followed the hypothesized theoretical distribution did not appear to be related. The correlational indices had relatively few "good fits" and those observed were nearly as likely to occur with samples of 15 as with samples of 120 or 60. If the distribution followed the hypothesized distribution, it tended to follow for all sample sizes for a given item, but not for all items.

2. Are the correlational type indices distributed as the product moment correlation coefficient?

The Kolmogorov-Smirnov goodness of fit test showed 256 of the 288 empirical sampling distributions were significantly different from the distribution of the product moment correlation coefficient corresponding to the population values of the indices. The results of the goodness of fit tests show quite clearly that the distribution of the product moment correlation coefficient was not an adequate model for the observed empirical distributions of the item discrimination indices. The detailed sources of the obtained discrepancies were not pursued in the present study; however, the bias, standard error, $\sqrt{\beta_1}$, and β_2 were reported for those who wish to pursue this point.

3. What are the relationships between the parameters of the item characteristic curve and the sampling properties of the indices?

The relationships among the mean value of a particular index and the corresponding item parameters were plotted in a three dimensional coordinate system having as its axes: X_{50}, β, and the index. If an item discrimination possesses desirable characteristics its mean value should form a surface which increases uniformly for all values of X_{50} as β increases. Only biserial r possessed this characteristic for all sample sizes. The remaining indices (r_f, r_c, and delta P) lost discriminating power as the item characteristic curve was moved away from the mid-point of the criterion scale, i.e., as the absolute value of X_{50} increased. When samples of 15 were employed, the decrease was substantial for r_f and delta P. Increasing sample size to 120 reduced the loss for r_c and r_f, but did not improve this characteristic for delta P.

The relationships of the standard deviations of the item discrimination indices to the item parameters X_{50} and β was not as consistent as those observed for the mean values of the indices.

When samples of size 15 were used, the standard deviations biserial r and Cook's index B increased, whereas those of $\sqrt{r_f}$ and P_d decreased as both X_{50} and β were simultaneously increased. The use of samples of size 120 reduced the magnitude of the observed standard deviations to approximately one third the value produced by samples of size 15. Biserial r and r_f continued to exhibit an increase in variability as the absolute values of X_{50} and β were both increased; however, the increase was much less pronounced than that observed in samples of size 15. When samples of 120 were employed, the standard deviation of delta P decreased slightly with increased magnitude of both X_{50} and β.

4. How do the sampling distributions of indices computed with internal and external criterion differ?

Although somewhat different sampling properties of the item discrimination indices were yielded by internal and external criterion scores, the apparent differences could be accounted for by the reduction in magnitude of the population value of the index. When the population values of the indices were similar, the observed sampling distributions based upon internal and external criterion were similar.

5. Under the conditions established, what are the relative merits of the selected indices?

A ranking of the indices based upon their empirical sampling characteristics would be as follows:

1. Maximum likelihood estimate of the item parameter β.
2. Biserial r

3. **Cook's index B**
4. **Product moment correlation coefficient estimated by the U-L 27% method.**
5. **Delta P**

It is somewhat difficult to make a clear-cut statement as to the desirability of an index, as they cannot be used without additional information. The correlational indices cannot be interpreted unless accompanied by the item difficulty and the item parameter β needs to be accompanied by the item parameter X_{50}. The empirical sampling properties of item difficulty agreed with those of the binomial distribution; however, virtually none of the empirical sampling distributions of the correlational type item discrimination indices could be shown to follow the distribution of the product moment correlation coefficient. The maximum likelihood estimates of the item parameters X_{50} and β produced a similar, though less restrictive, set of characteristics. The empirical sampling distributions of β followed those of the normal distribution under nearly all conditions. The obtained sampling distributions of X_{50} occasionally followed the normal distribution. Only one of the pair had very desirable properties which was also the case with item difficulty and the correlational type indices. From a practical and a theoretical point of view, however, the over-all sampling characteristics of β were uniformly desirable, while the sampling properties of X_{50} became deviant only when the item lacked discrimination, thus, the better the item the better the sampling characteristics of the item parameters. The converse held true for the item difficulty-correlational index approach, item difficulty had relatively good sampling characteristics, whereas the correlational indices had uniformly undesirable characteristics. The item parameters X_{50} and β provide a much firmer conceptual basis for the description and interpretation of an item's characteristics than does the item difficulty-correlational index approach. In addition, the item parameter frame work enables one to continue the development of modern mental test theory as established by Lawley [31] employing the powerful characteristics of the normal distribution. If one is unable to divorce themselves from the conceptual constraints of the item difficulty-correlational index methods of item analysis, the empirical results would indicate the use of biserial r and item difficulty.

In the past, the maximum likelihood estimation of the item parameters X_{50} and β has been economically impractical, but the present widespread availability of digital computers makes this method of item analysis no more difficult than any other.

An ancillary part of the present study was the investigation of the empirical sampling characteristics of the Hoyt reliability coefficient [22]. The Hoyt reliability coefficient appears to have desirable sampling characteristics for sample sizes, number of items, and the single population value (.906) employed in the present study. The bias of the mean value of r_h was small for samples as small as size 15 and was negligible for samples of size 120. In addition, the range of obtained values was small (.87 to .94) when samples of 60 and 120 were employed. In the present case, the results indicated that if the population value of the reliability coefficient was .906, a reasonably close estimate could be obtained for samples of size 60 or larger. When samples of 120 were employed, the observed distribution of the Hoyt reliability coefficient and the distribution of the product moment correlation coefficient corresponding to the square root of the reliability coefficient adjusted for a small bias did not differ significantly. The question which needs to be reconciled with the present data, however, is the effect of the number of items. In the present case, the distribution of the product moment correlation coefficient for the sample size was employed, ignoring the number of items. The Hoyt method is a function of both the number of items and individuals, thus, leaving one in a quandry as to the effect of items. The central F distribution did not prove to be a good model for the distribution of the Hoyt reliability coefficient; however, a possibility exists that the non-central F could be the appropriate model. On the basis of the present results, the distribution of the product moment correlation coefficient corresponding to the $\sqrt{r_h}$ appears to be a more likely model when samples as large as 120 are employed.

Several procedural problems arose which need to be investigated in some detail. Considerable time and effort was expended to obtain the third and fourth moments of the empirical distributions and compute Pearson's coefficients of skewness, $\sqrt{\beta_1}$ and kurtosis, β_2. Although Pearson [37] reported the theoretical values of $\sqrt{\beta_1}$ and β_2 for the various distributions of the product moment correlation coefficient, they appear to be of no value. The author was unable to locate a statistical technique whereby one could test whether the empirical values of $\sqrt{\beta_1}$ and β_2 differed significantly from the theoretical values; except in the case of the normal distribution. Either the author doesn't know where to look, or the distributions of $\sqrt{\beta_1}$ and β_2 for non-normal situations are unknown.

A second problem arose when attempting to compute the theoretical values of $\sqrt{\beta_1}$ and β_2 of the distribution of the product moment correlation coefficient using Pearson's formulas [37, page CLXII] and single precision floating point arithmetic. The computer results agreed with those given by Pearson [37] only in the case of the first two moments. None of the values of $\sqrt{\beta_1}$ and β_2 agreed with the tabled values. Desk calculations employing ten decimal places also could not produce the tabled values using the data given in the examples by Pearson [37]. The difficulties encountered seem to indicate the need for the use of multiple precision arithmetic when implementing Pearson's formulas [37].

The random sampling and subsequent analysis of the item data generated approximately 15,000 punched cards each containing 14 units of information for a total of 210,000 units. Although computer errors are relatively rare occurrences, the sheer mass of data made the occurrences of some errors inevitable, thus data checking procedures were mandatory. In projects of this type one should plan on the occurrence of errors and institute checking procedures at an early date. The preliminary analyses employing samples of size 15 exhibited the majority of the malfunctions occurring in the system, hence, served admirably to debug the total data processing procedure. The majority of errors detected were attributed directly to the card punching equipment, whether computer directed or not. Despite the cautions given above, the functioning of the total computer data generating and processing system was extremely gratifying. Only the Kolmogorov-Smirnov goodness of fit tests utilizing computer generated observed and theoretical distributions were performed by hand.

Taken as a composite, the results of the present study show clearly the inadvisability of attempting to perform item analyses based upon small samples. The single item discrimination index based upon samples of size 15, 30, or 60, has little or no value and those based upon samples of 120, with the possible exception of the item parameter β, are only indicative in nature. It remains for future research to establish lower bounds for sample size when the present item analysis techniques are employed.

REFERENCES

1. Adams, J. F. "Effect of Non-Normally Distributed Criterion Scores on an Item Analysis Technique," *Educational and Psychological Measurement*, Vol. 20, No. 2, Summer 1960, pp. 317-320.

2. Baker, F. B. "Univac Scientific Computer Program for Test Scoring and Item Analysis," *Behavioral Science*, Vol. 4, CPA 1, 1959, pp. 254-255.

3. Butler, J. M. "A Ratio for Estimating the Reliability of Test Scores," *Journal of Educational Psychology*, Vol. 33, 1942, pp. 391-395.

4. Carter, H. D. "How Reliable are the Common Measures of Difficulty and Validity of Objective Test Items," *Journal of Psychology*, Vol. 13, 1942, pp. 31-39.

5. Clark, E. L. "Item Difficulties Based Upon End Segments," *Journal of Educational Psychology* 48, November, 1957, pp. 457-459.

6. Conrad, H. S. "Characteristics and Uses of Item Analysis Data," *Psychological Monograph, General and Applied*, 1948, 62 No. 8, Whole No. 295.

7. Cook, W. W. *The Measurement of General Spelling Ability Involving Controlled Comparisons Between Techniques*, University of Iowa Studies in Education, Vol. VI, No. 6, 1932.

8. David, F. N. *Tables of Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples*, Issued by Biometrika Office University College, London, Cambridge; The University Press, 1938.

9. Davis, F. B. "Item Analysis Data," *Harvard Educational Papers No. 2*, Cambridge, Massachusetts, Graduate School of Education, Harvard University.

10. Fan C. T. "Item Analysis Table," Princeton, New Jersey, Educational Testing Service, 1952.

11. Fan, C. T. "Note on Construction of an Item Analysis Table for the High-Low 27% Group Method," *Psychometrika*, Vol. 19, 1954.

12. Finney, D. J. "The Application of the Probit Analysis to the Results on Mental Tests," *Psychometrika*, Vol. 19, 1944, pp. 31-39.

13. Finney, D. J. *Probit Analysis.* Cambridge: The University Press, 1952.

14. Flanagan, J. C. "General Considerations in the Selection of Test Items and a Short Method of Estimating the Product Moment Correlation Coefficient from Data at the Tails of the Distribution," *Journal of Educational Psychology*, Vol. 30, 1939, pp. 674-680.

15. Fowler, H. M. "An Application of the Ferguson Method of Computing Item Conformity and Person Conformity," *Journal of Experimental Education*, Vol. 22, 1954, pp. 237-246.

16. Garwood, F. "The Application of Maximum Likelihood to Dosage-Mortality Curves," *Biometrika*, Vol. 32, 1941, pp. 46-58.

17. Gibbons, C. C. "The Predictive Value of the Most Valid Items of an Examination," *Journal of Educational Psychology*, 1940, pp. 616-621.

18. Guilford, J. P. *Psychometric Methods.* New York and London: McGraw and Hill Book Company, 1936.

19. Guilliksen, H. *Theory of Mental Tests.* New York: John Wiley and Sons; London: Chapman and Hall, 1950.

20. Holloway, C. Jr., and Capp, W. B. "IBM 704 Program for Fisher's F-Distribution" *Behavioral Science*, 6, 1961, p. 371.

21. Hoyt, C. J. "Test Reliability Estimated by Analysis of Variance," *Psychometrika* VI, 1941, pp. 153-160.

22. Hoyt, C. J. and MacEachern, D. G. Unpublished Research, University of Minnesota, 1958.

23. Jackson, R. W. B. "Reliability of Mental Tests," *British Journal Psychology*, 29, 1939, pp. 267-287.

24. Johnson, A. P. "Notes on a Suggested Index of Item Validity," *Journal of Educational Psychology*, 1951, 42, pp. 499-504.

25. Johnson, P. O. *Statistical Methods in Research*, New York: Prentice-Hall, Inc., 1949.

26. Jones, R. A. and Michael, W. B. "An Empirical Study of Systemative Errors in Estimates of Item Difficulty Obtained from Use of Upper and Lower 27% Criterion Groups," *Educational and Psychological Measurement* 17, 31, 135, Spring, 1957.

27. Kelley, T. L. The Kelley Statistical Tables. New York: Macmillan Company, 1938.

28. Kelley, T. L. "The Selection of Upper and Lower Groups for the Validation of Test Items," Journal of Educational Psychology, Vol. 30, 1939, pp. 17-24.

29. Kuang, H. P. "A Critical Evaluation of the Relative Efficiency of Three Techniques in Item Analysis," Educational and Psychological Measurement, Vol. 12, 1952, pp. 248-266.

30. Kuder, G. F. and Richardson, M. W. "The Theory of the Estimation of Test Reliability," Psychometrika, Vol. 2, No. 3, September, 1937, pp. 151-160.

31. Lawley, D. N. "On Problems Connected With Item Selection and Test Construction," Royal Society Edinburgh Proceedings, Vol. 61A, 1943, pp. 273-287.

32. Long, J. A. and Sandiford, P. The Validation of Test Items, Bulletin No. 3, The Department of Educational Research Toronto: The University of Toronto Press, 1935.

33. Lord, Frederic, A Theory of Test Scores. Psychometric Monographs No. 7, 1952.

34. Maritz, J. A. "Estimation of the Correlation Coefficient in the Case of a Bivariate Normal Distribution when one of the Variables is Dichotomized," Psychometrika, Vol. 18, 1953, pp. 97-110.

35. McNamara, W. J. and Dunlap, J. W. "A Graphical Method for Computing the Standard Error of Biserial R, Journal of Experimental Education, Vol. 2, 1934, pp. 274-277.

36. Pearson, Karl, "On a New Method of Determining Correlation Between a Measured Character A and a Character B, Of Which Only the Percentages of Cases Wherein B Exceeds (Or falls short) of a Given Intensity is Recorded for Each Grade of A," Biometrika, Vol. VII, 1909, pp. 96-105.

37. Pearson, Karl, Editor, Tables for Statisticians and Biometricians, Part II, First Edition, Issued by the Biometrika Office University College, London: Cambridge: The University Press, 1931.

38. Remington Rand, Engineering Research Associates Division, The ERA 1103 Computer System, St. Paul, Minnesota, 1951.

39. Ross, C. C., and Stanley, J. C. Measurement in Today's Schools. Englewood Cliffs, New Jersey: Prentice Hall, Inc. 1954.

40. Siegel, Sidney, Nonparametric Statistics for the Behavioral Sciences, New York, Toronto, and London: McGraw-Hill Book Company, 1956.

41. Sopher, H. E. "On the Probable Error of the Bi-Serial Expression of the Correlation Coefficient," Biometrika, Vol. X, 1914.

42. Special Projects Office, Bureau of Naval Weapons, Department of the Navy. Pert Instruction Manual and Systems and Procedures for the Program Evaluation System, Washington, D. C., 1960.

43. Vernon, P. E. "Indices of Item Consistency and Validity," British Journal of Psychology, Statistical Section, Vol. 2, 1941, pp. 152-166.

44. The Staff of the Computation Laboratory, Tables of the Cumulative Binomial Probability Distribution, Harvard University Press, Cambridge, Massachusetts, 1955.

45. Walker, Helen M., and Lev, J. Statistical Inference, New York: Henry Holt and Company, 1953.

PREFACE TO APPENDIX A

The empirical frequency distributions of the selected item discrimination indices obtained in the present study are presented in this appendix. The results for each index are in the following order: Items one through nine based upon an internal criterion, items one through nine based upon an external criterion.

Figure 24. Frequency distribution of item difficulty for the nine items as obtained from random samples of size 15, 30, 60, 120.

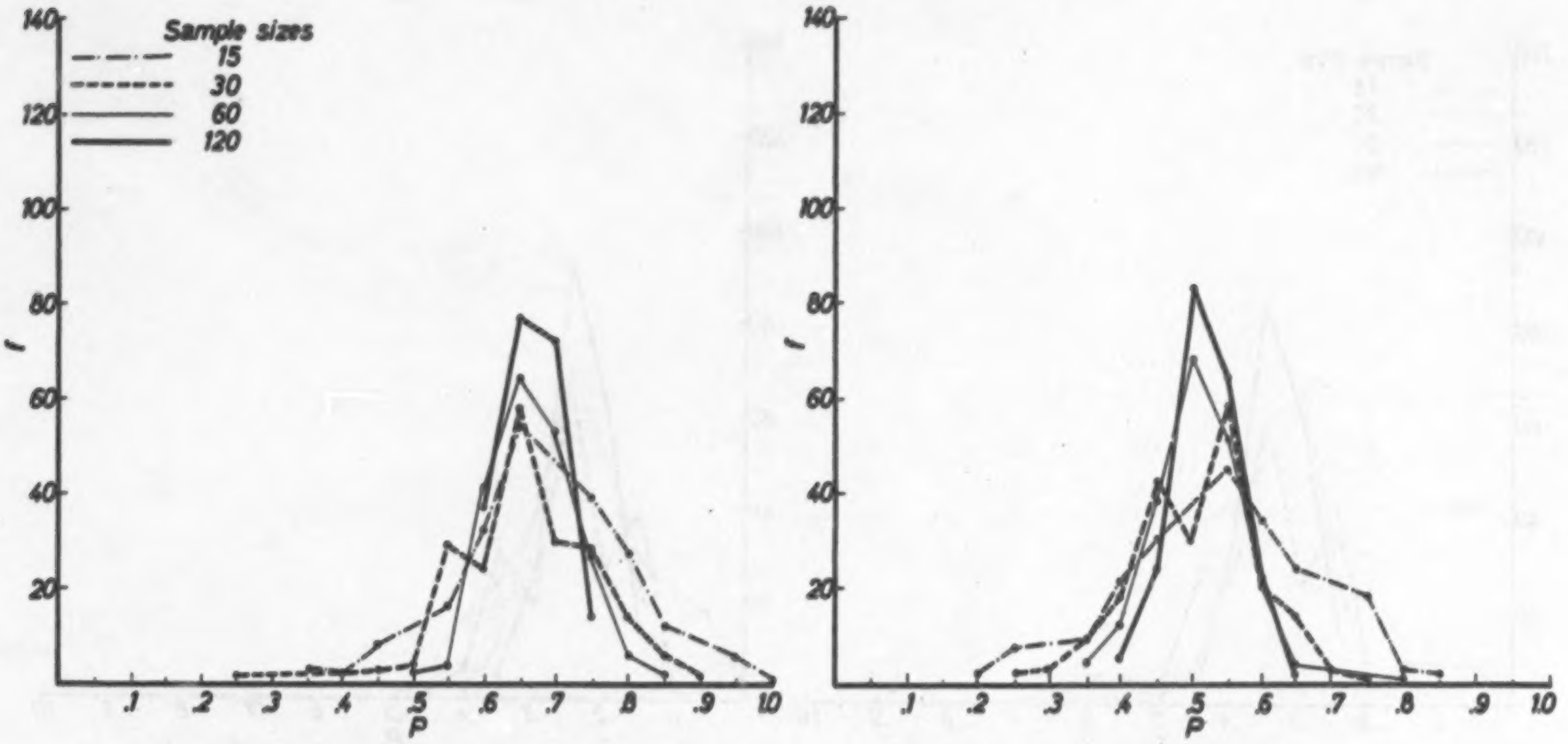


Figure 24a. Item 1, P=.692

Figure 24b. Item 2, P=.506

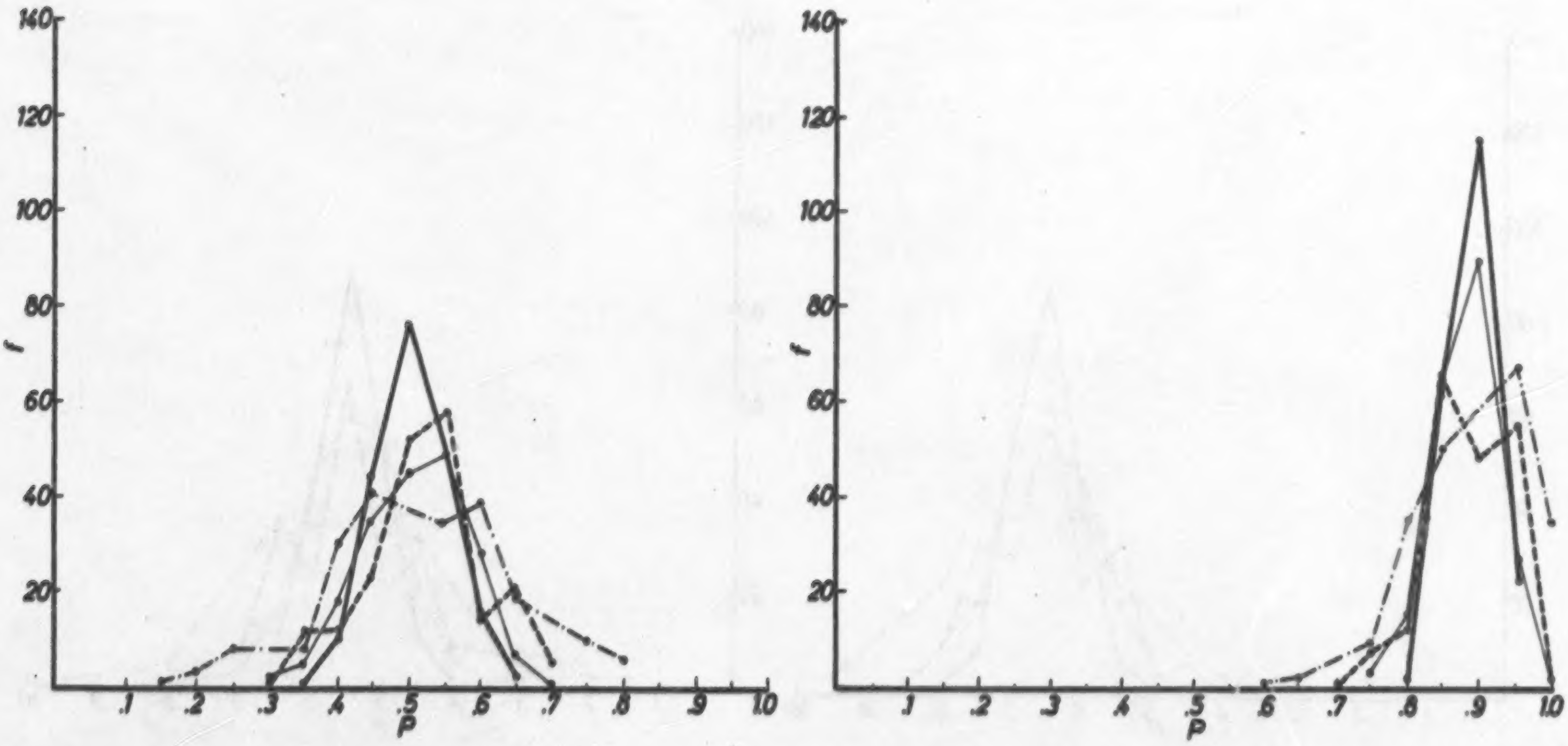


Figure 24c. Item 3, P=.497

Figure 24d. Item 4, P=.888

Figure 24. continued.

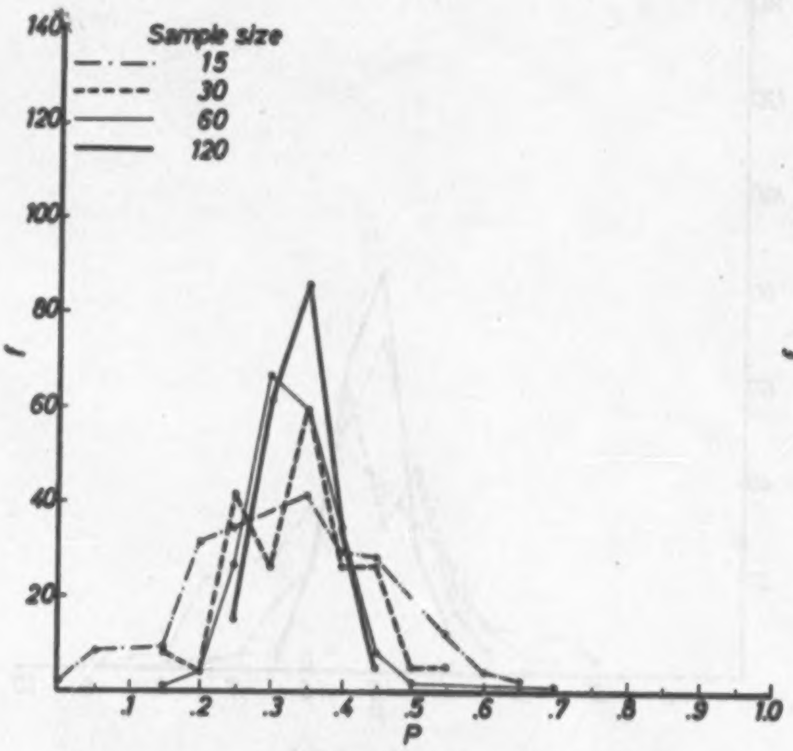


Figure 24e. Item 5, P=.332

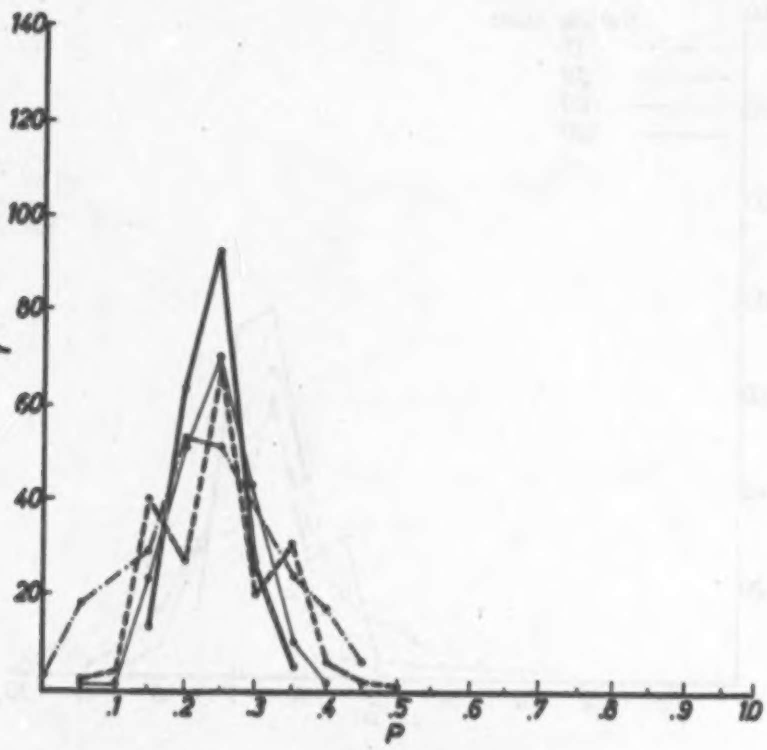


Figure 24f. Item 6, P=.229

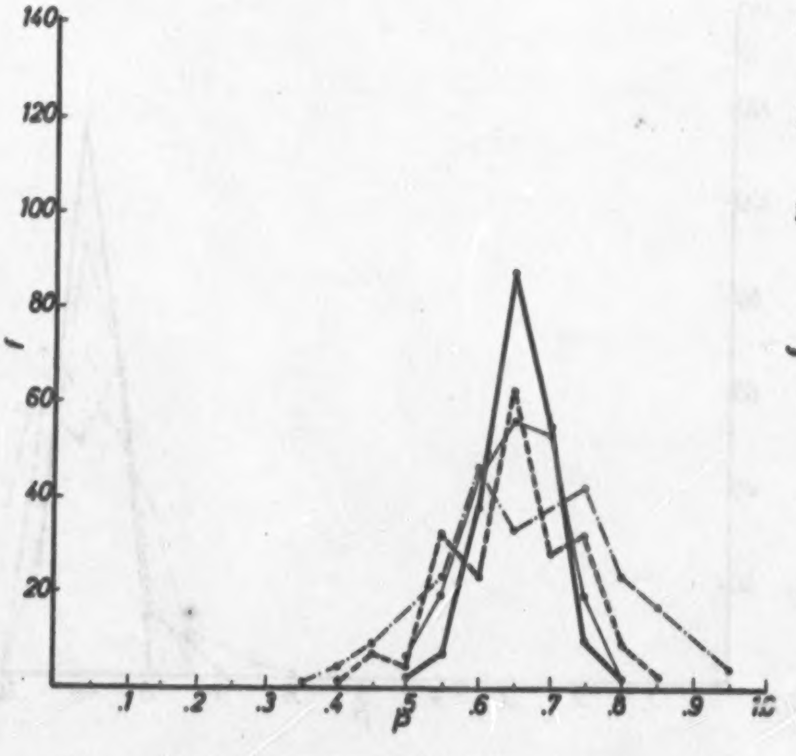


Figure 24g. Item 7, P=.659

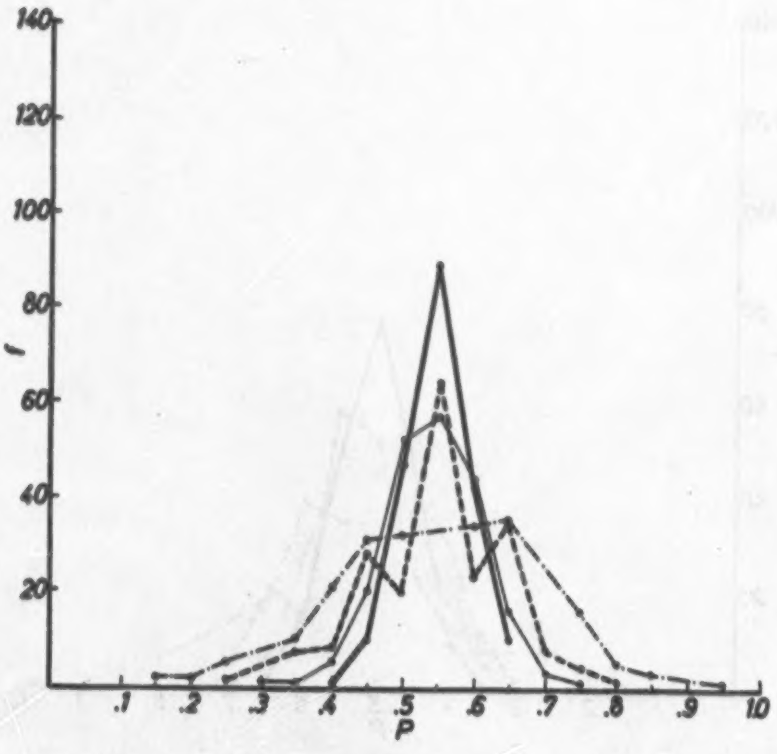


Figure 24h. Item 8, P=.534

Figure 24 continued.

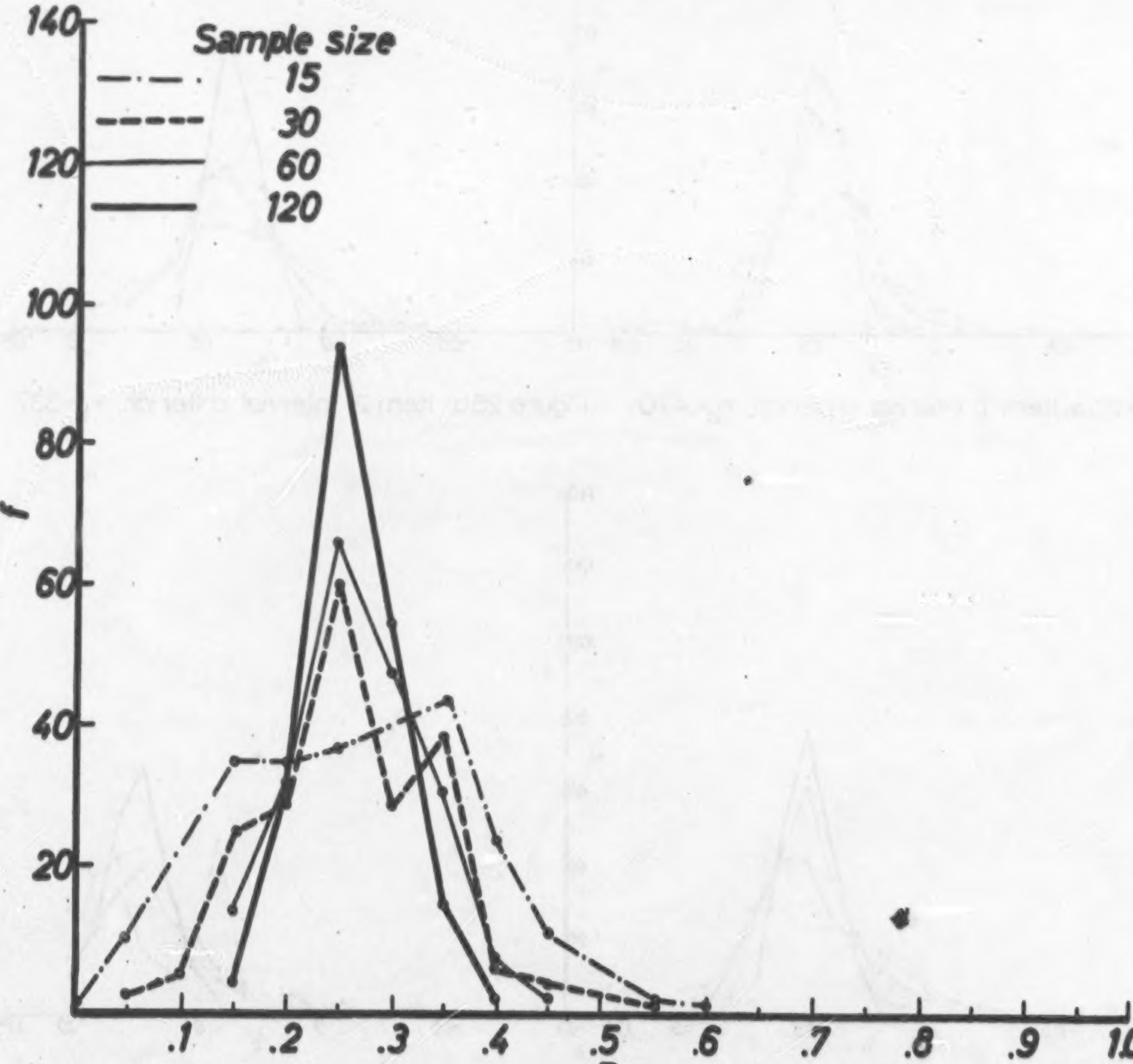


Figure 24 i. Item 9, P=.265

Figure 25. Frequency distribution of biserial r for the nine items as obtained from random samples of size 15, 30, 60, 120, internal and external criterion.

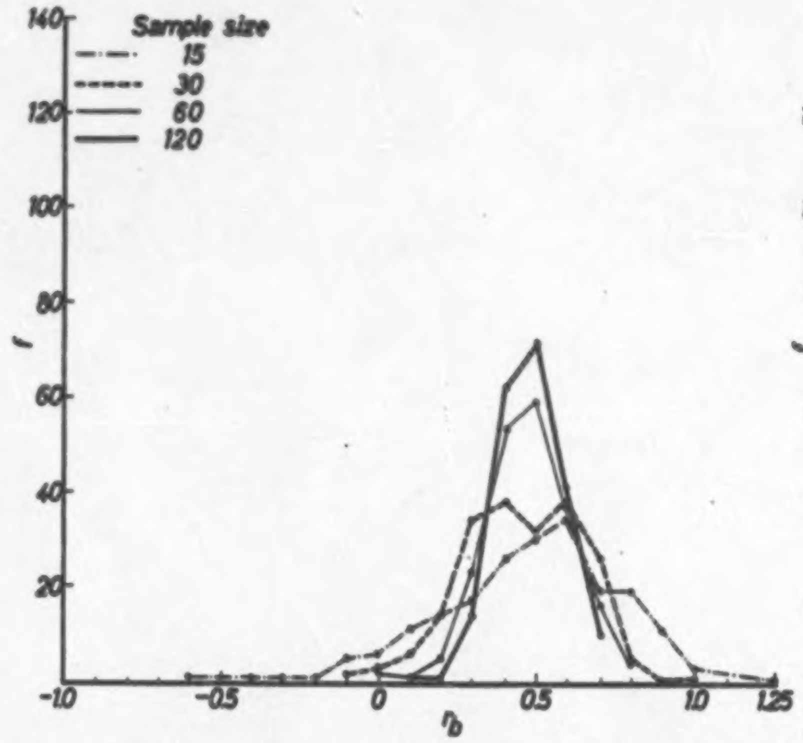


Figure 25a. Item 1, internal criterion, r_b=.470

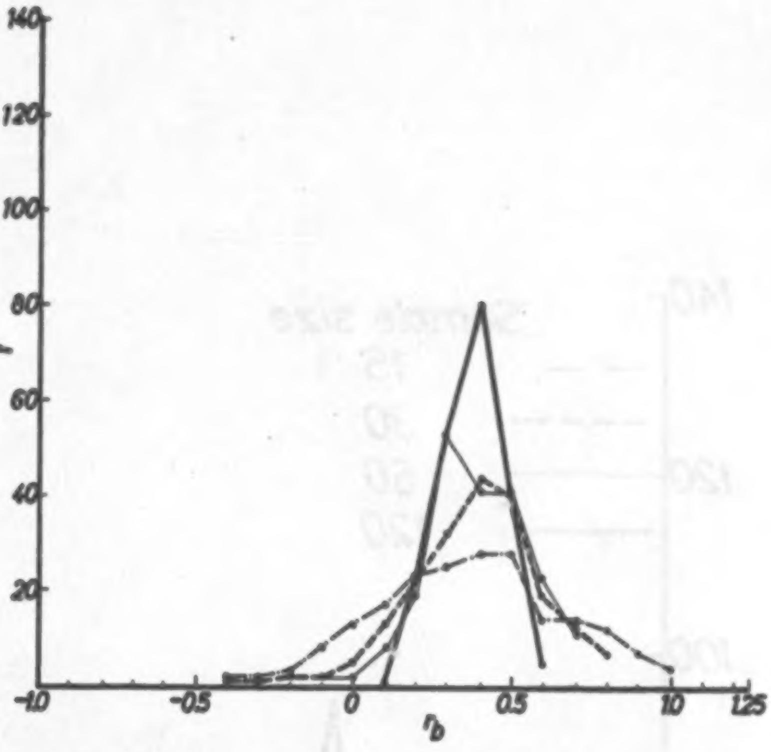


Figure 25b. Item 2, internal criterion, r_b=.337

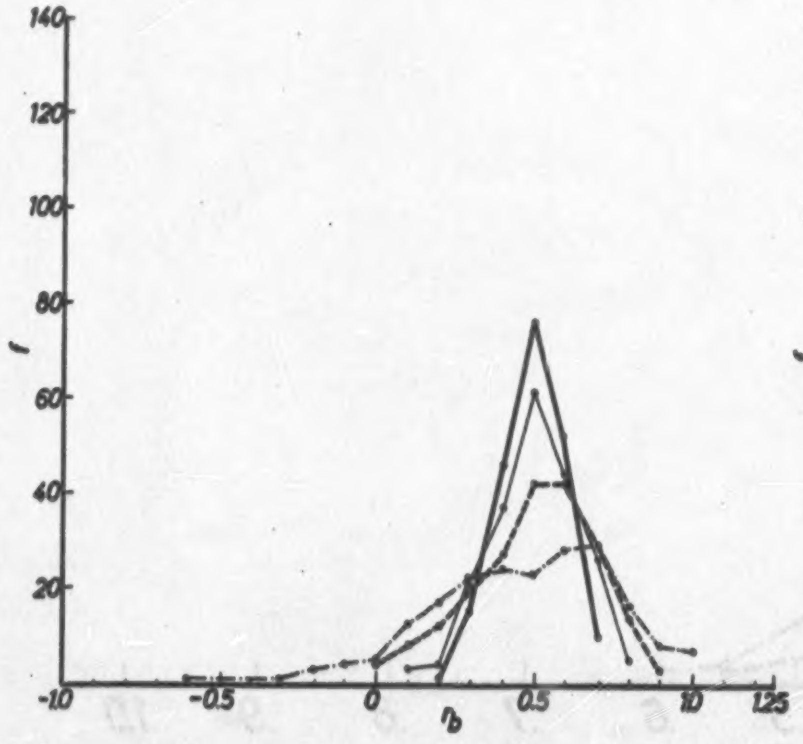


Figure 25c. Item 3, internal criterion, r_b=.524

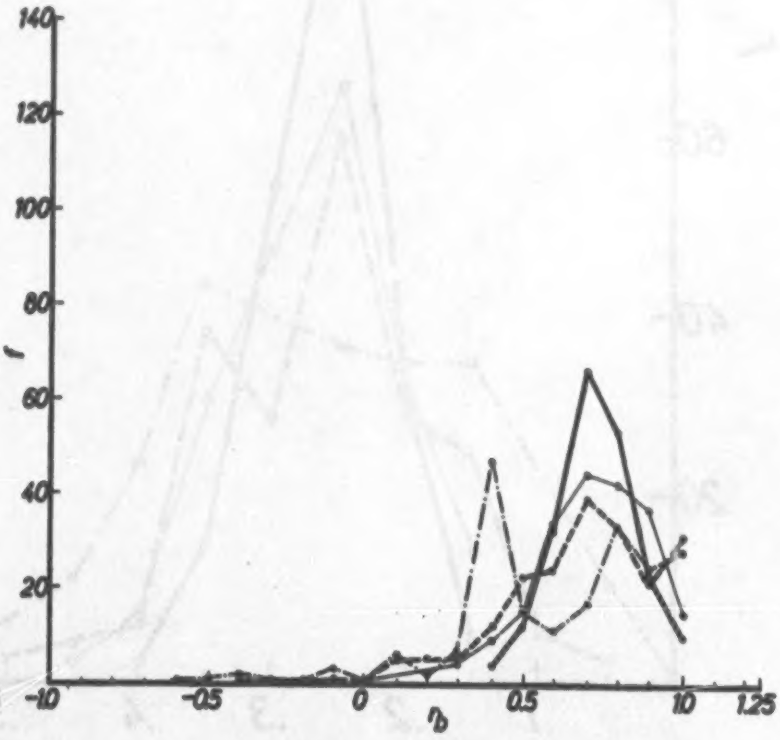


Figure 25d. Item 4, internal criterion, r_b=.711

Figure 25. continued.

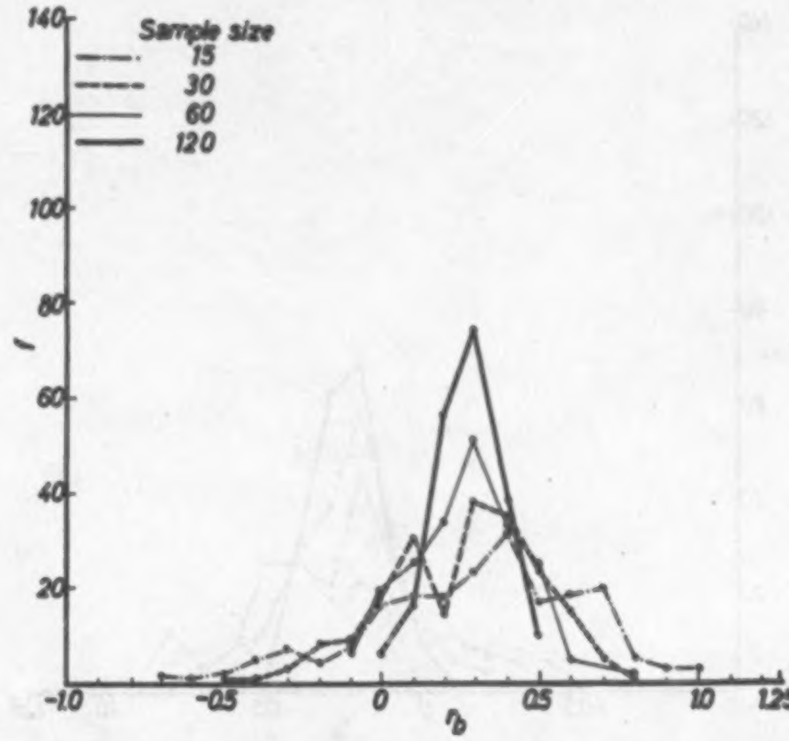


Figure 25e. Item 5, internal criterion, r_b=.231

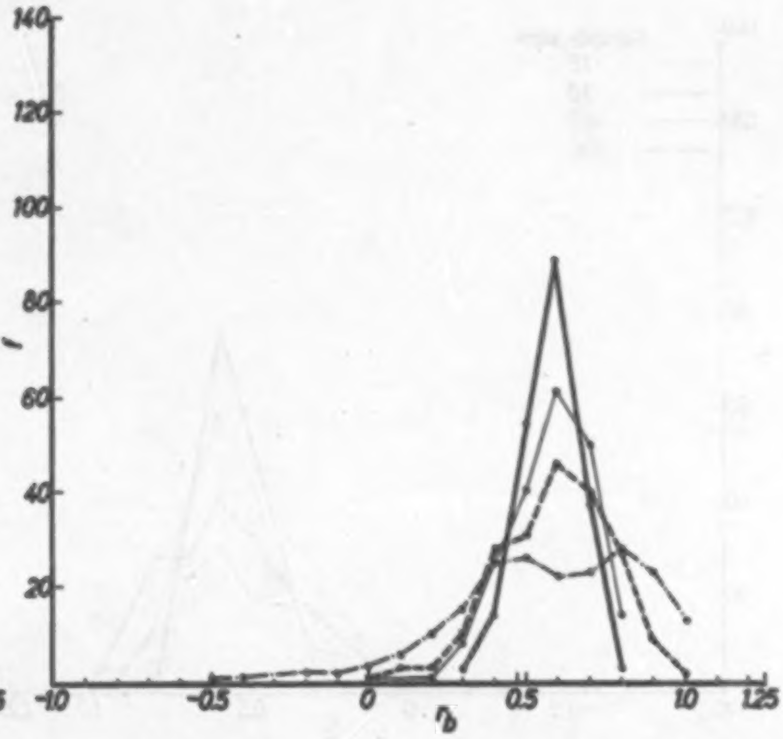


Figure 25f. Item 6, internal criterion, r_b=.580

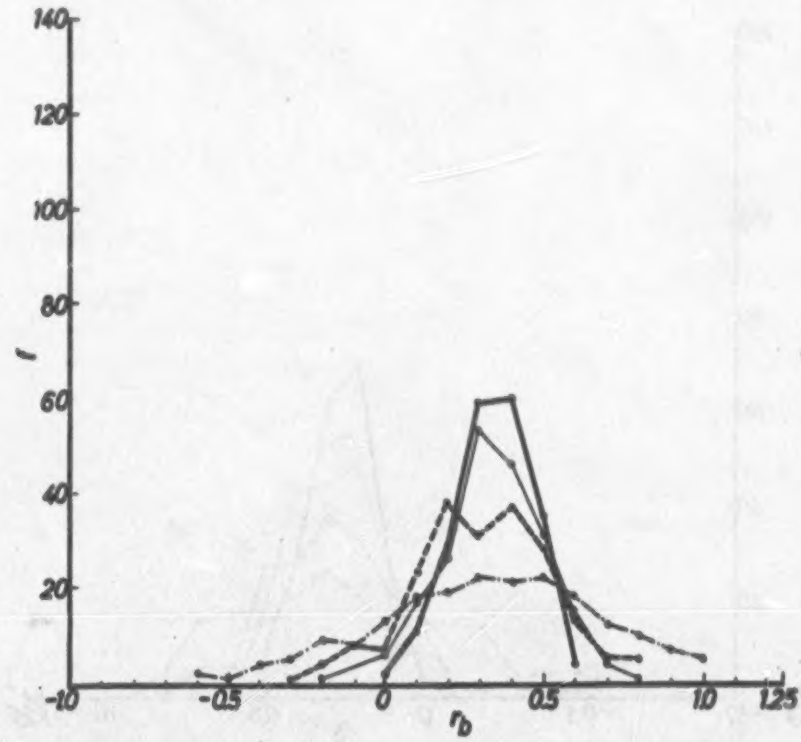


Figure 25g. Item 7, internal criterion, r_b=.340

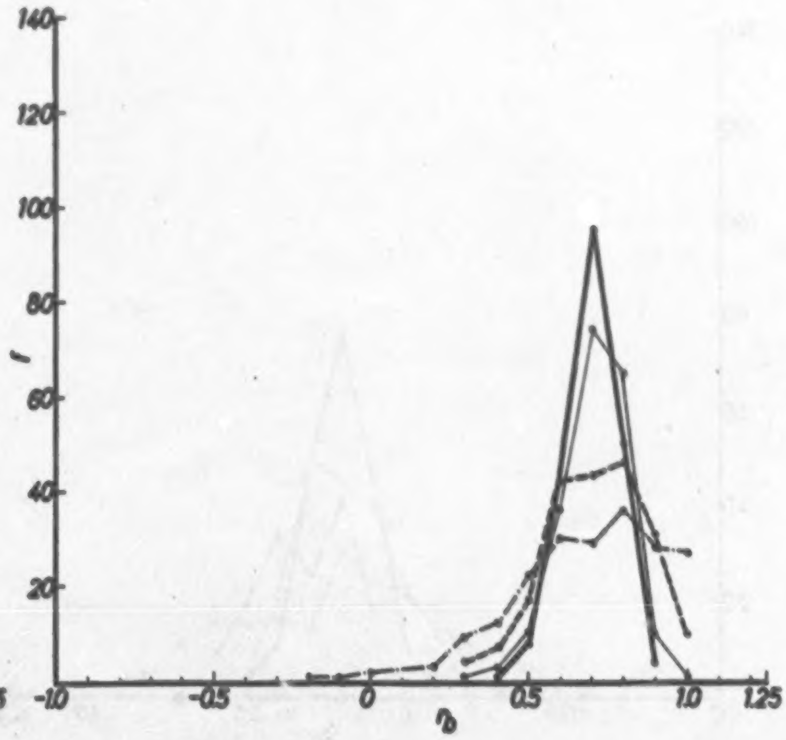


Figure 25h. Item 8, internal criterion, r_b=.695

Figure 25. continued

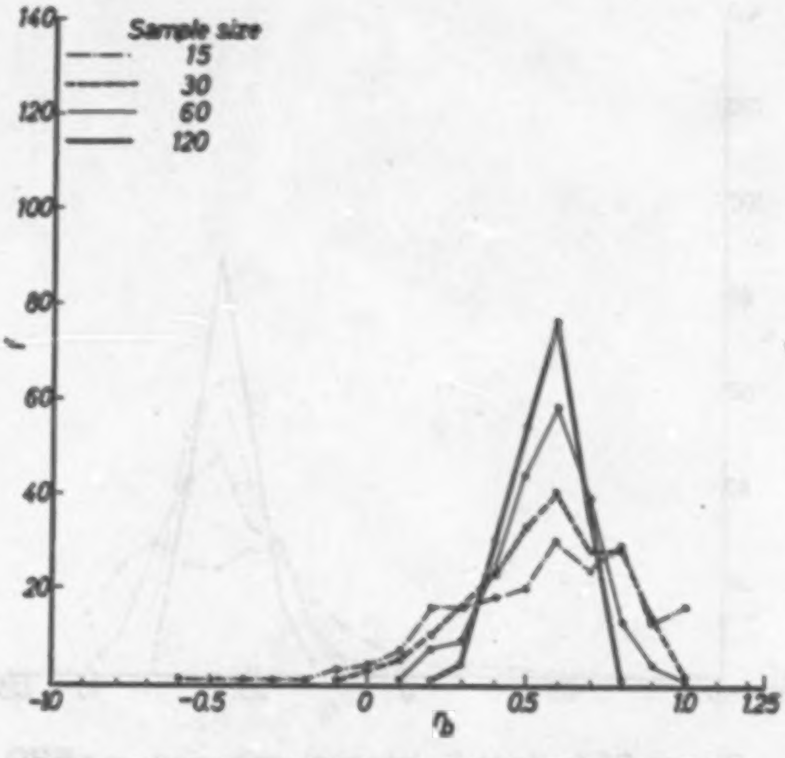


Figure 25i. Item 9, internal criterion, r_b=.550

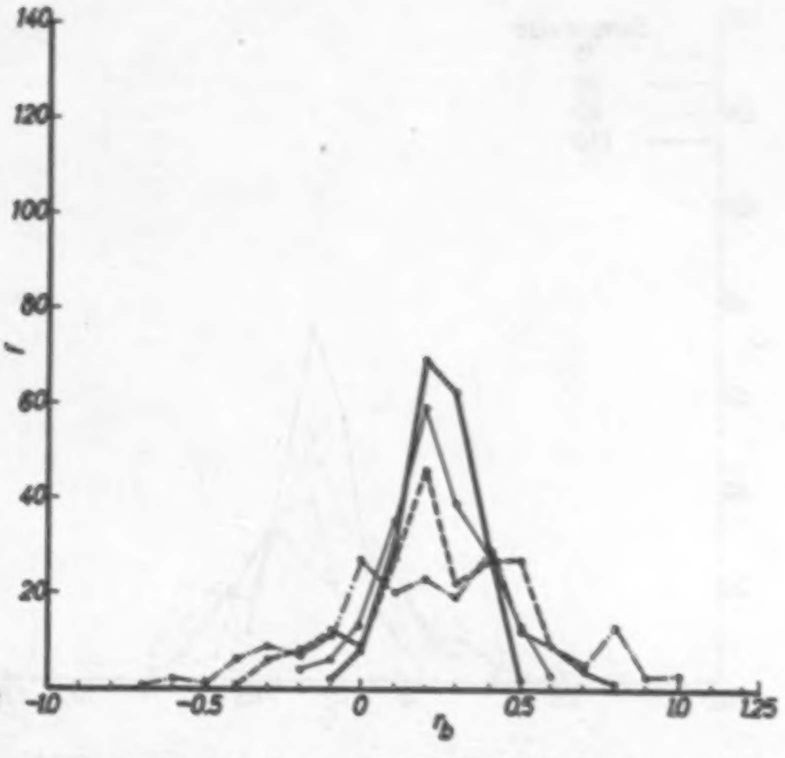


Figure 25j. Item 1, external criterion, r_b=.197

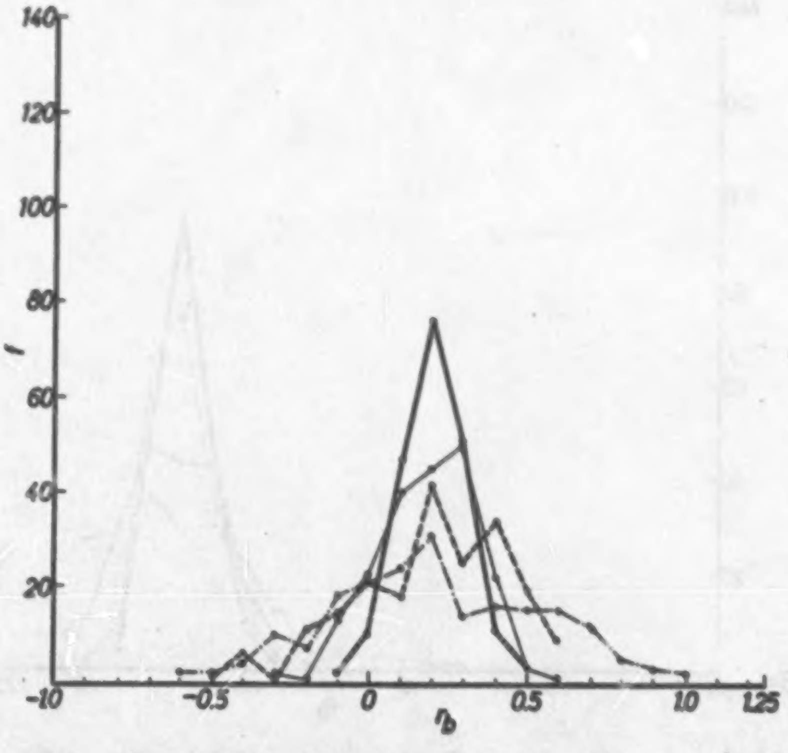


Figure 25k. Item 2, external criterion, r_b=.200

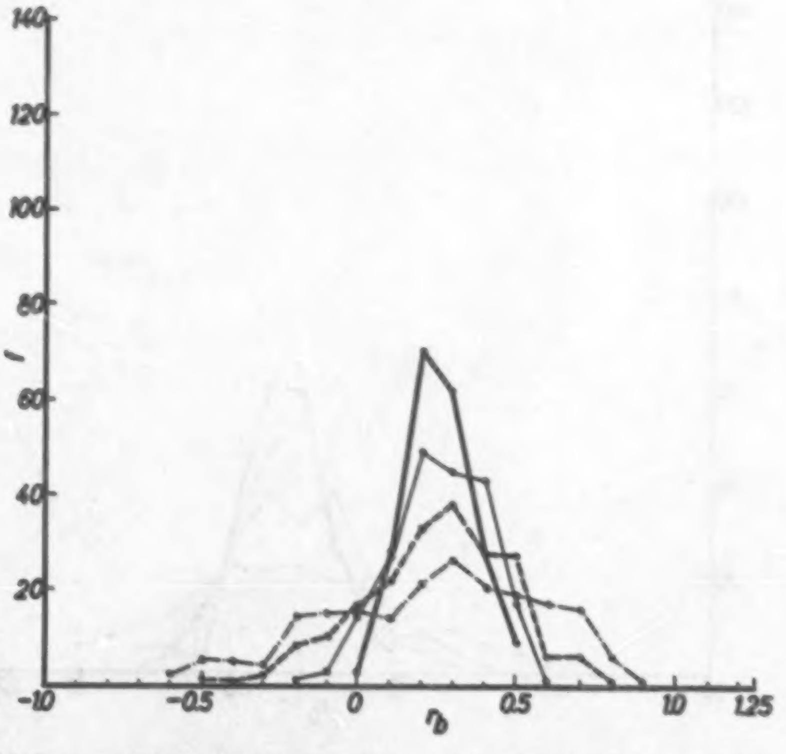


Figure 25l. Item 3, external criterion, r_b=.228

Figure 25. continued.

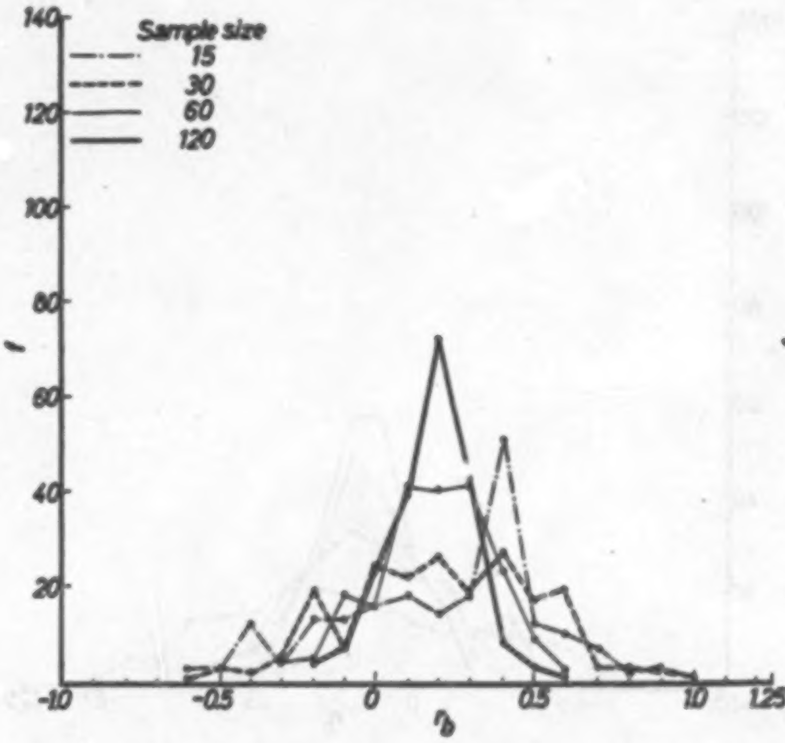


Figure 25m. Item 4, external criterion, r_b=.189

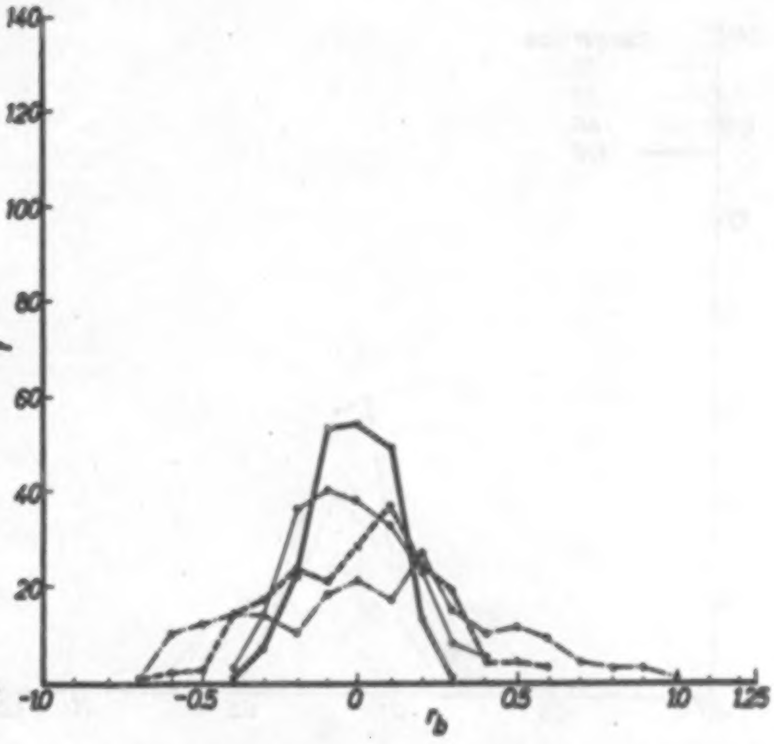


Figure 25n. Item 5, external criterion, r_b=-.013

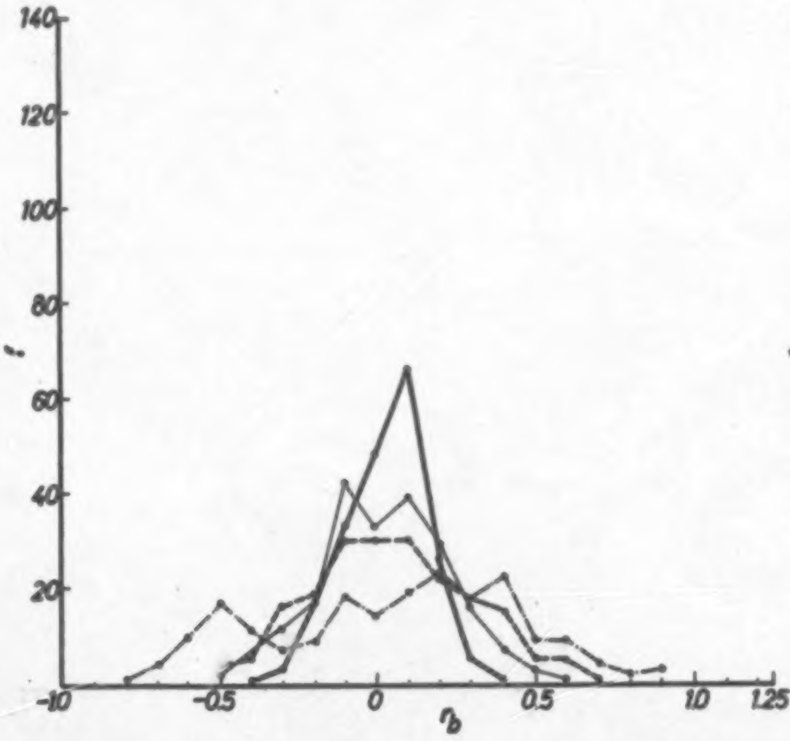


Figure 25o. Item 6, external criterion, r_b=.068

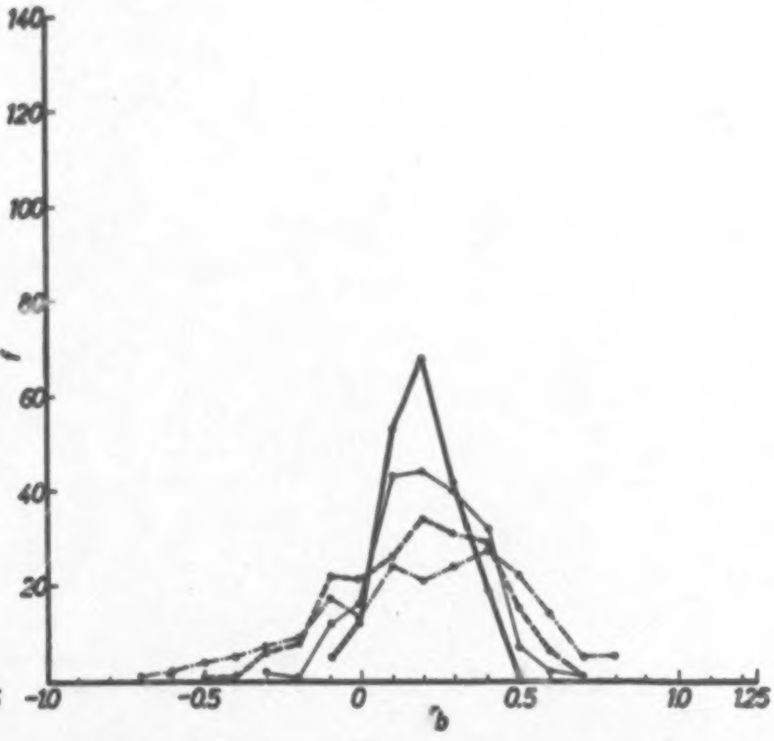


Figure 25p. Item 7, external criterion, r_b=.186

Figure 25. continued.

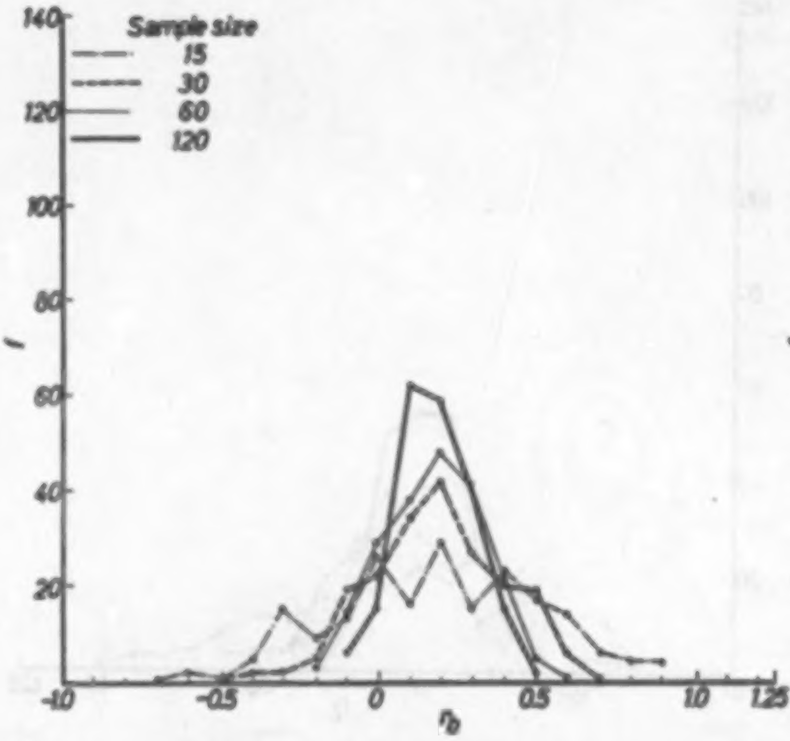


Figure 25q. Item 8, external criterion, r_b=.142

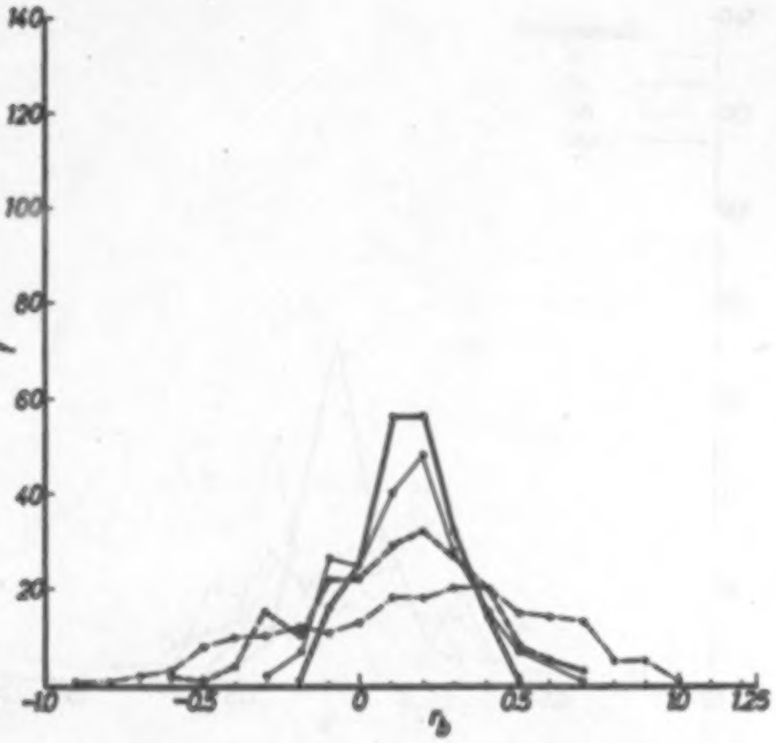


Figure 25r. Item 9, external criterion, r_b=.149

Figure 26. Frequency distribution of r_f for the nine items as obtained from random samples of size 15, 30, 60, 120, internal and external criterion.

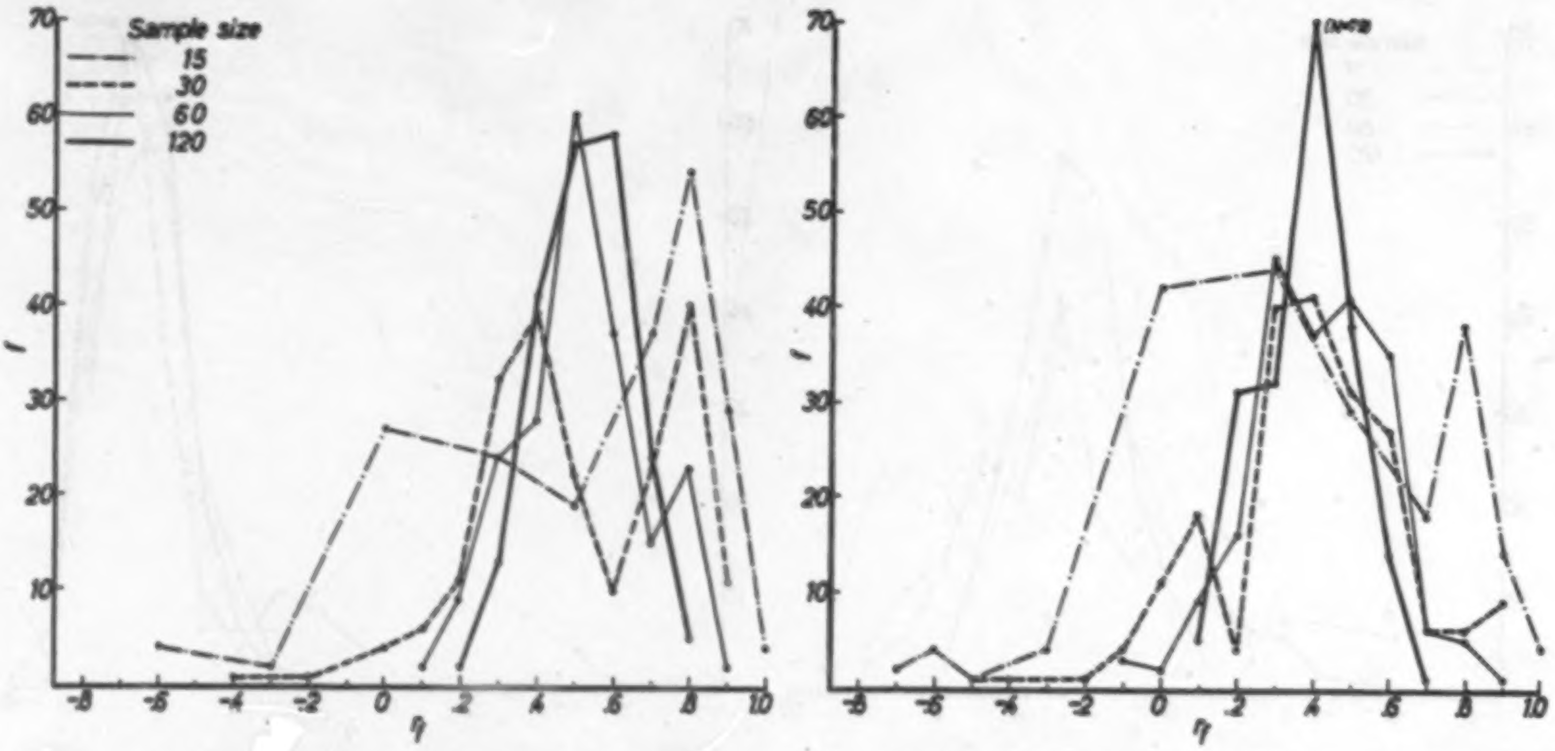


Figure 26a. Item 1, internal criterion, η=.500

Figure 26b. Item 2, internal criterion, η=.340

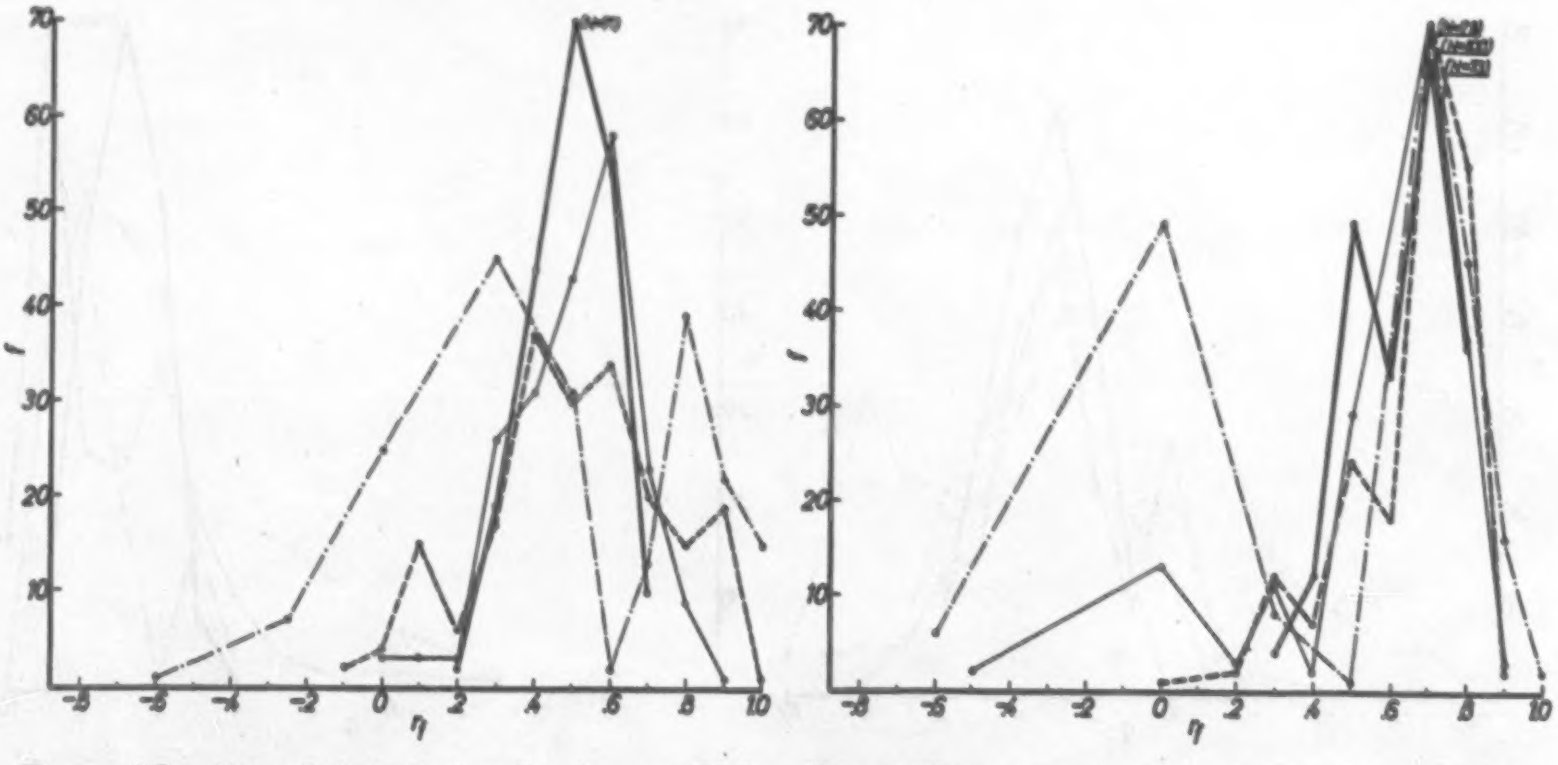


Figure 26c. Item 3, internal criterion, η=.520

Figure 26d. Item 4, internal criterion, η=.630

Figure 26. continued.

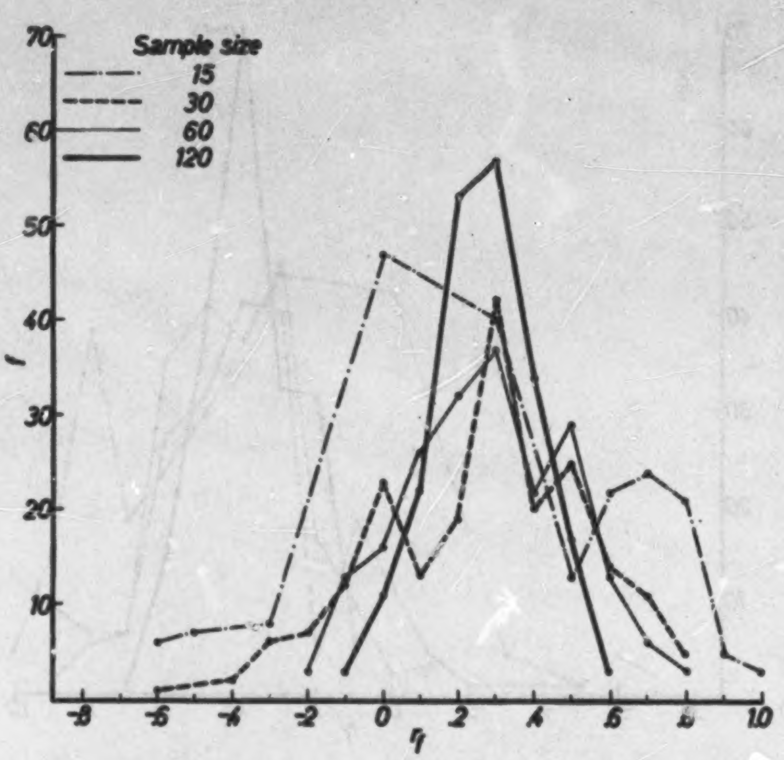


Figure 26e. Item 5, internal criterion, r_I=.230

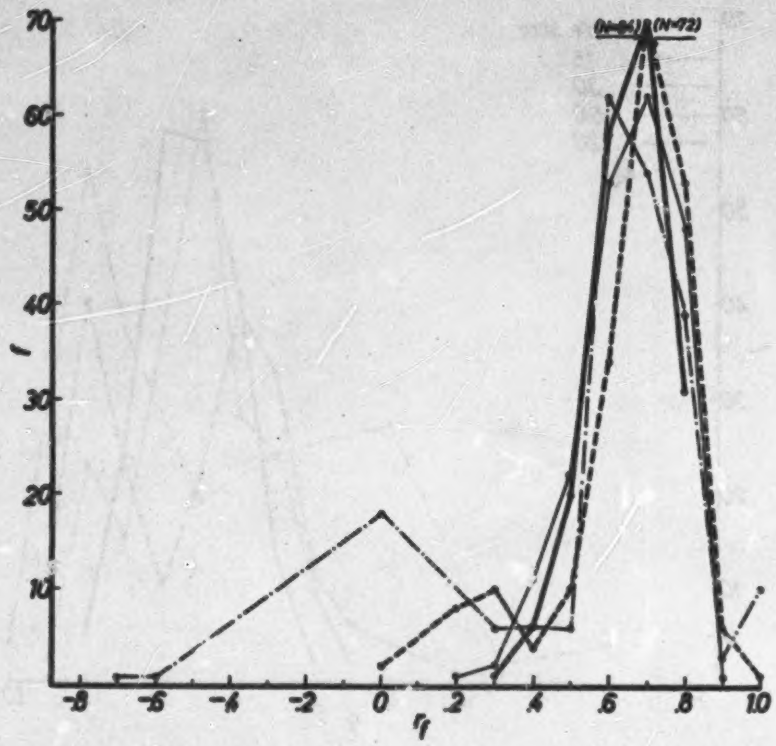


Figure 26f. Item 6, internal criterion, r_I=.630

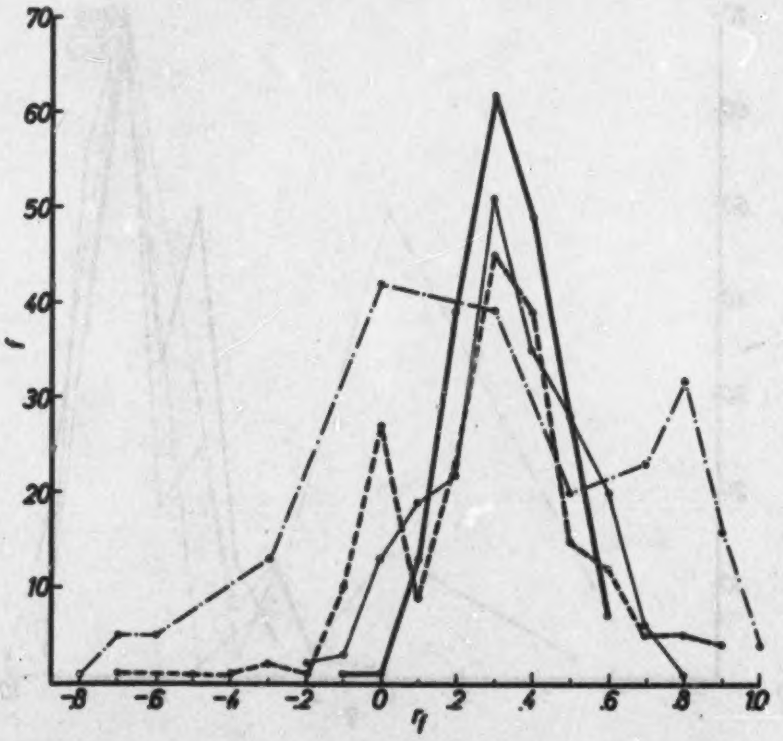


Figure 26g. Item 7, internal criterion, r_I=.300

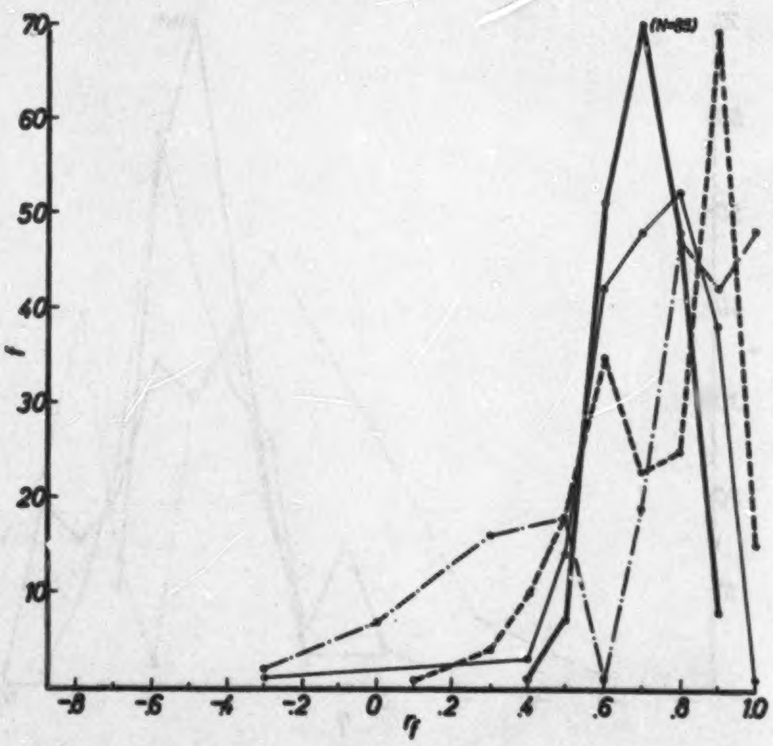


Figure 26h. Item 8, internal criterion, r_I=.670

3

Figure 26. continued.

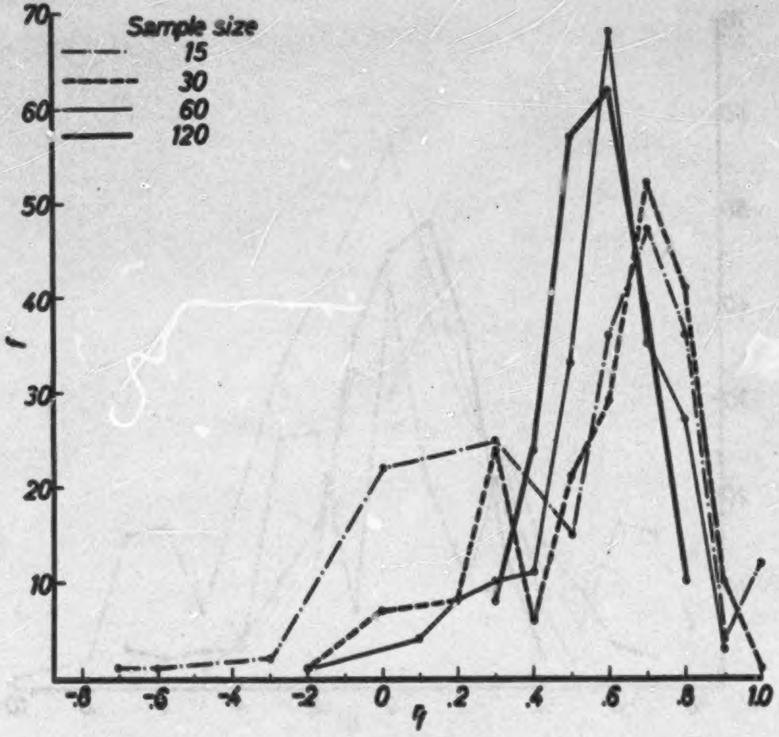


Figure 26i. Item 9, internal criterion, r_t=.550

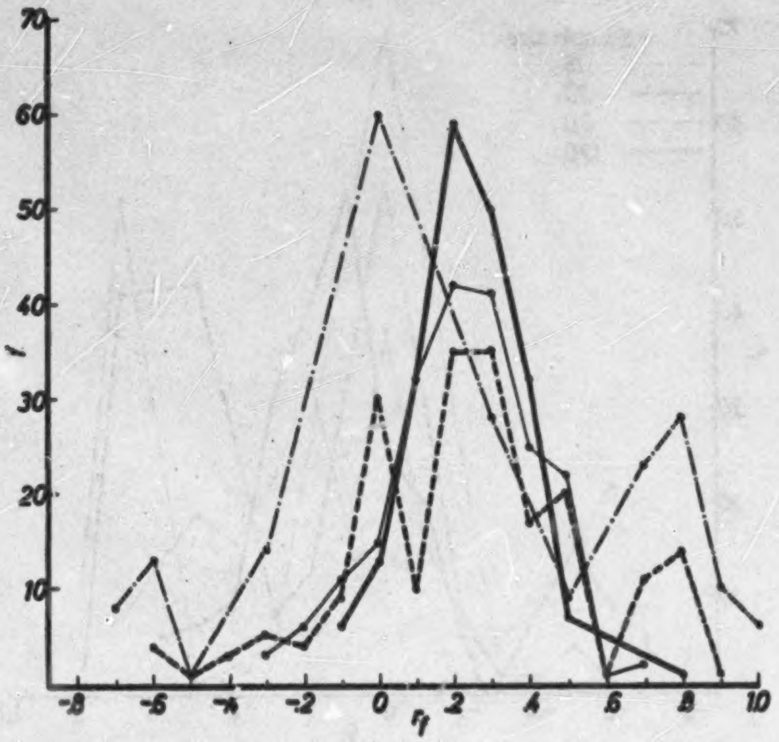


Figure 26j. Item 1, external criterion, r_t=.200

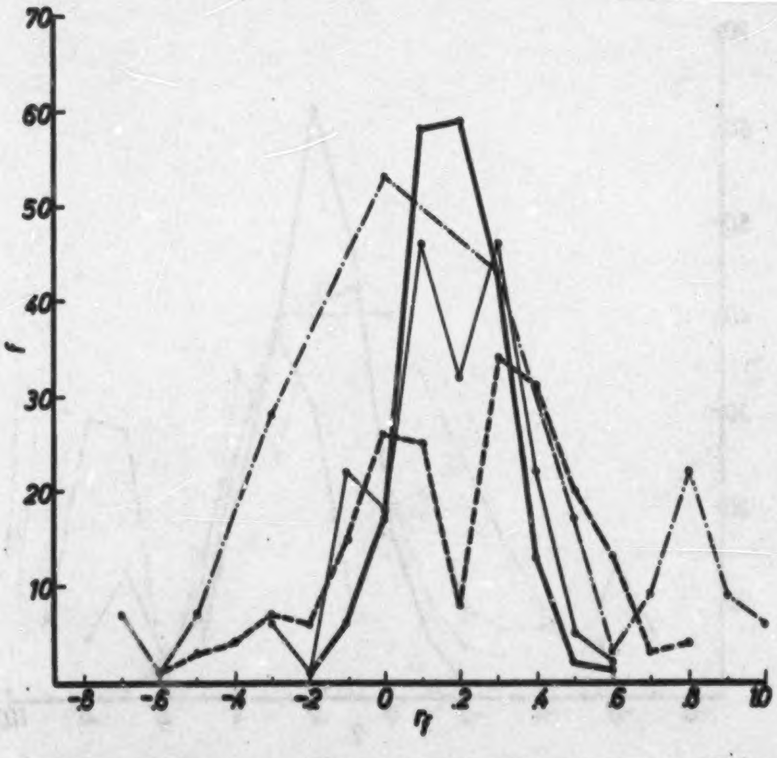


Figure 26k. Item 2, external criterion, r_t=.170

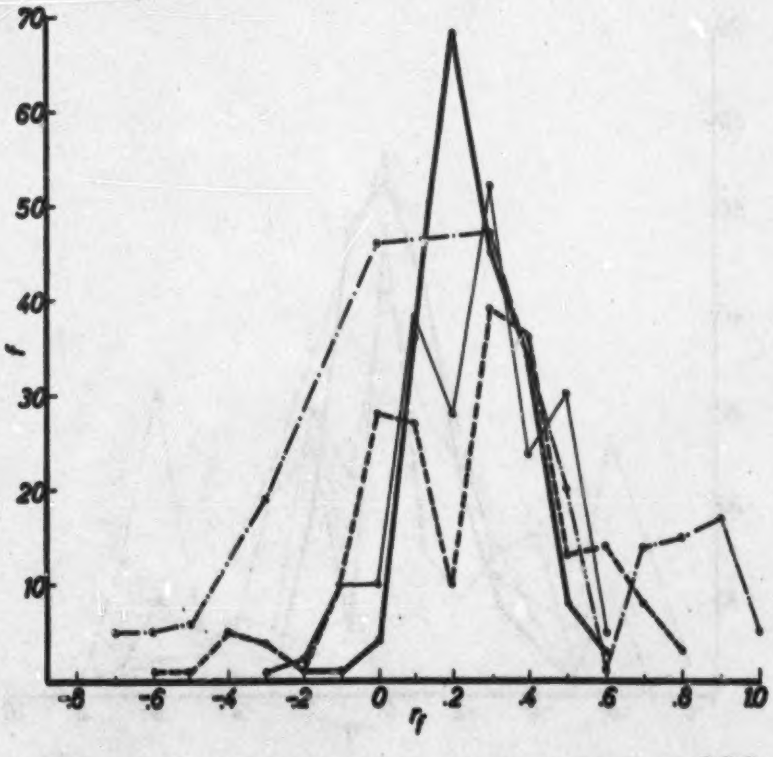


Figure 26l. Item 3, external criterion, r_t=.230

Figure 26. continued.

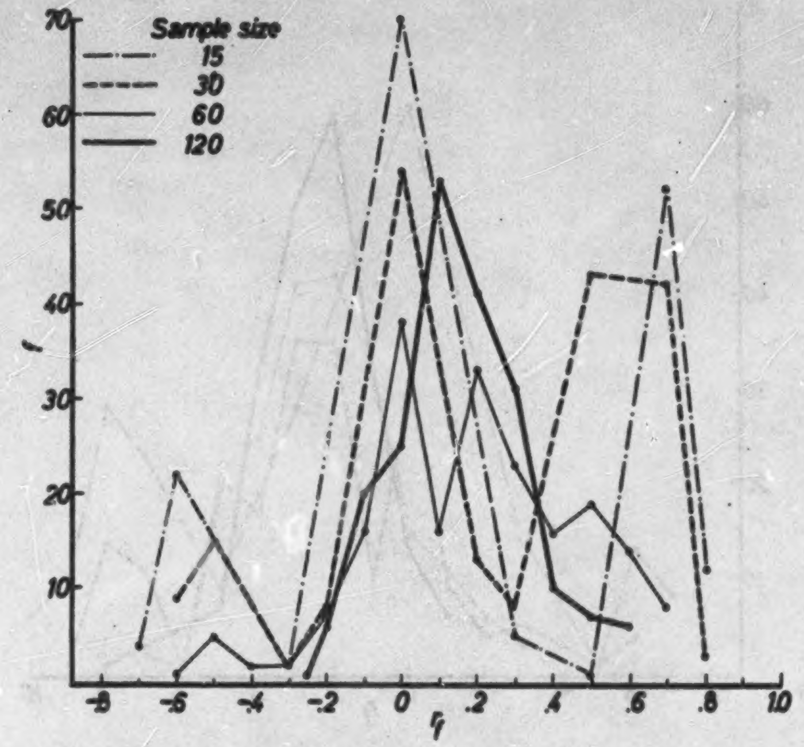


Figure 26m. Item 4, external criterion, r_f=.180

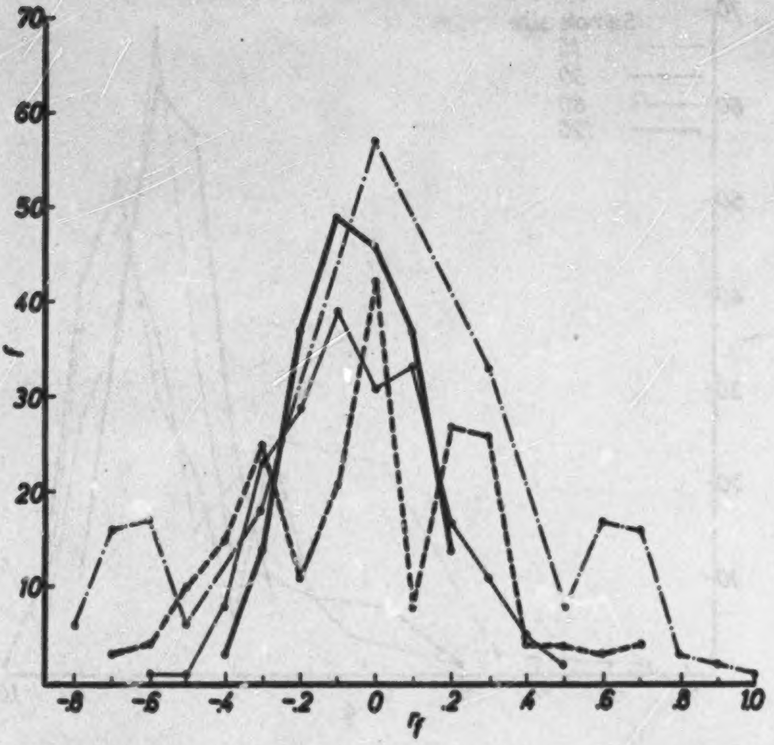


Figure 26n. Item 5, external criterion, r_f=-.040

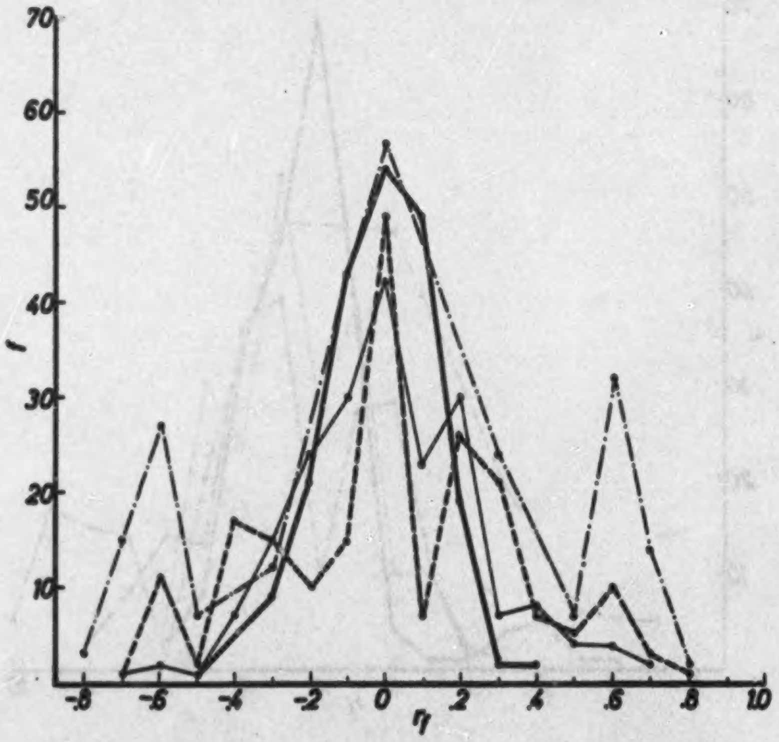


Figure 26o. Item 6, external criterion, r_f=0.000

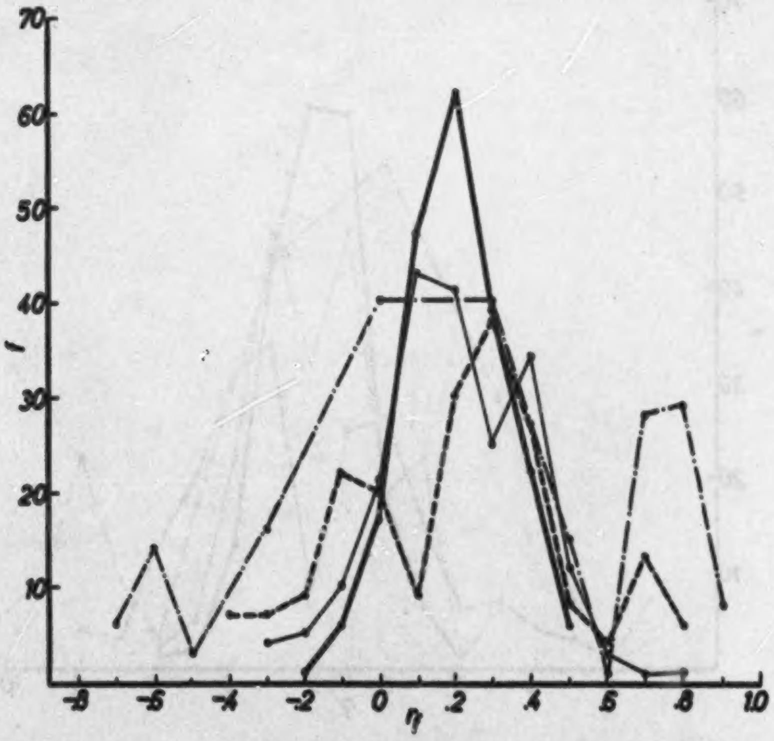


Figure 26p. Item 7, external criterion, r_f=.200

Figure 26. continued.

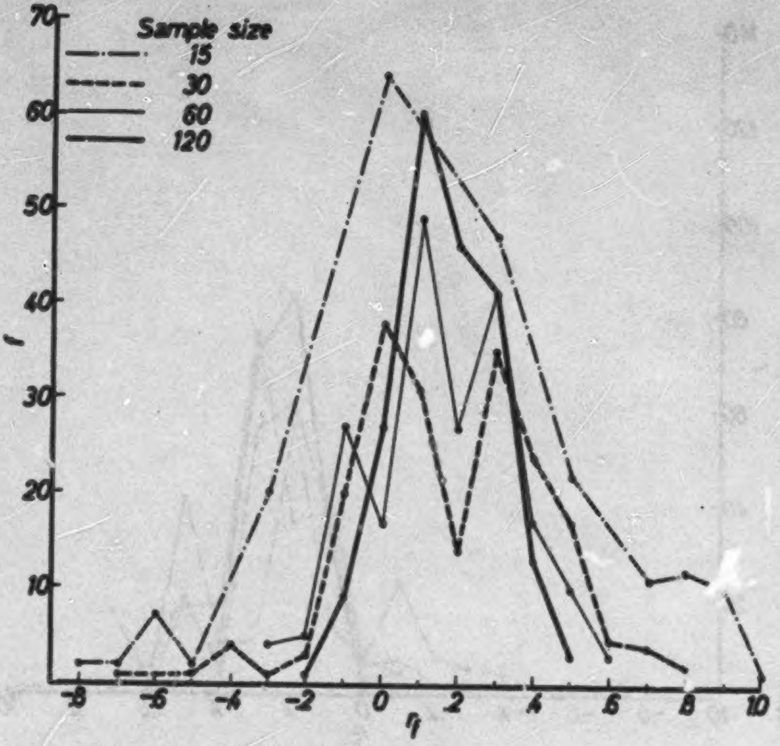


Figure 26q. Item 8, external criterion, r_i=.120

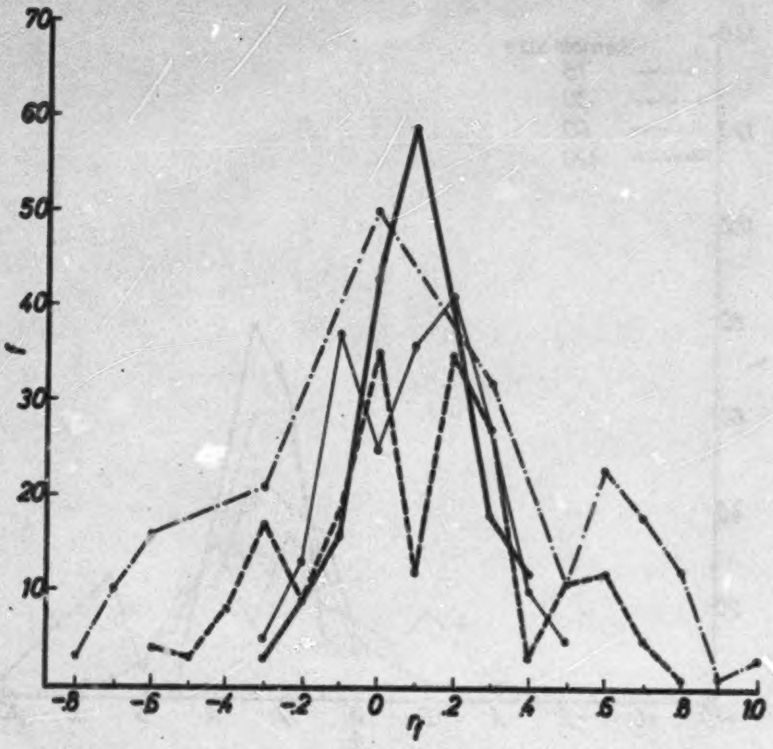


Figure 26r. Item 9, external criterion, r_i=.110

Figure 27. Frequency distribution of r_c for the nine items as obtained from random samples of size 15, 30, 60, 120, internal and external criterion.

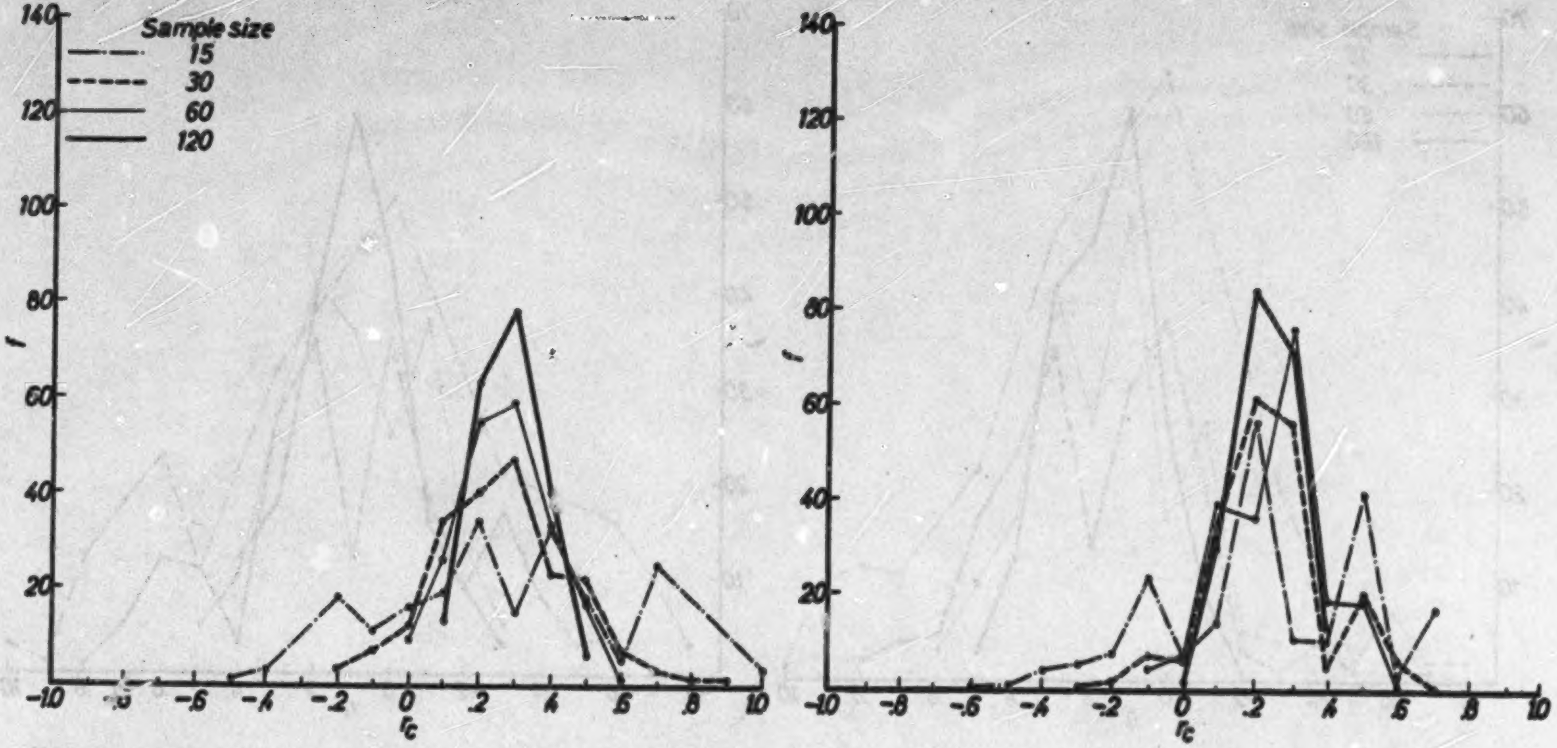


Figure 27a. Item 1, internal criterion, r_c=.259

Figure 27b. Item 2, internal criterion, r_c=.202

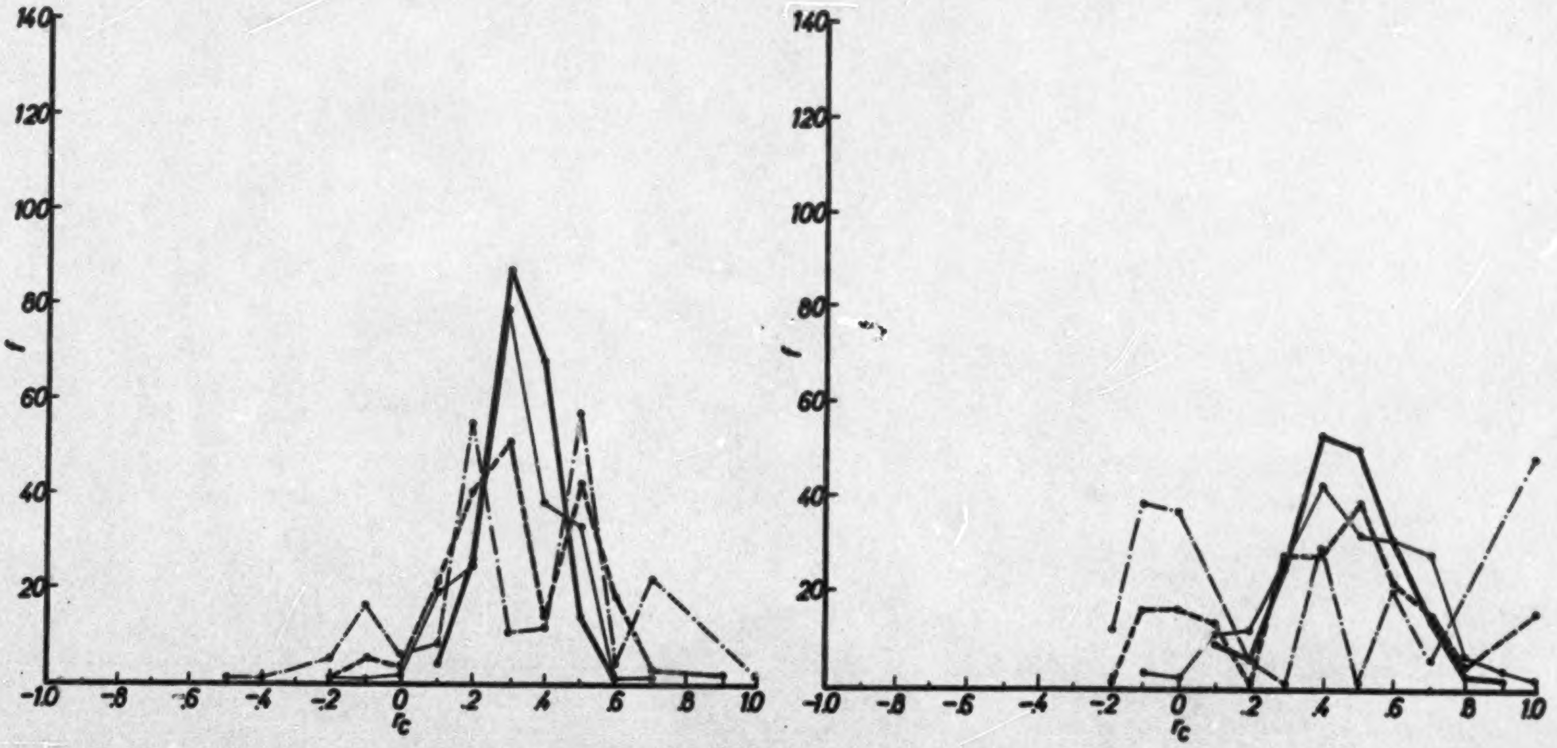


Figure 27c. Item 3, internal criterion, r_c=.347

Figure 27d. Item 4, internal criterion, r_c=.463

Figure 27. continued.

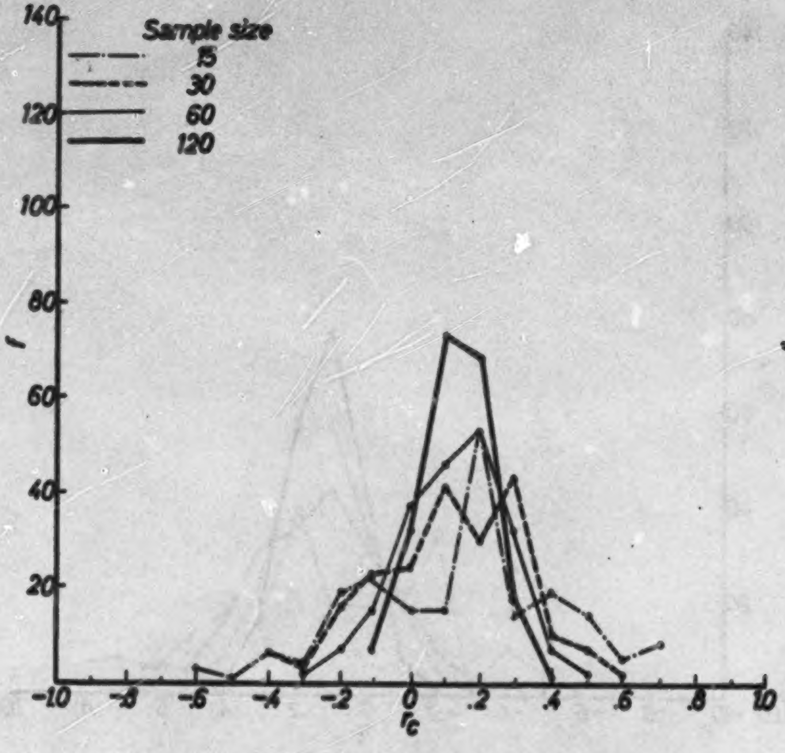


Figure 27e. Item 5, internal criterion, r_c=.119

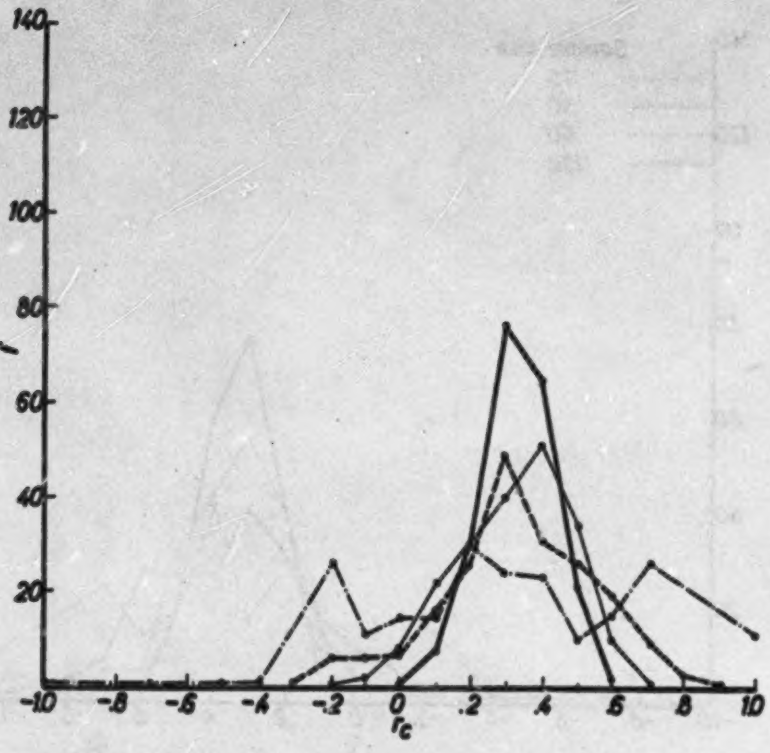


Figure 27f. Item 6, internal criterion, r_c=.363

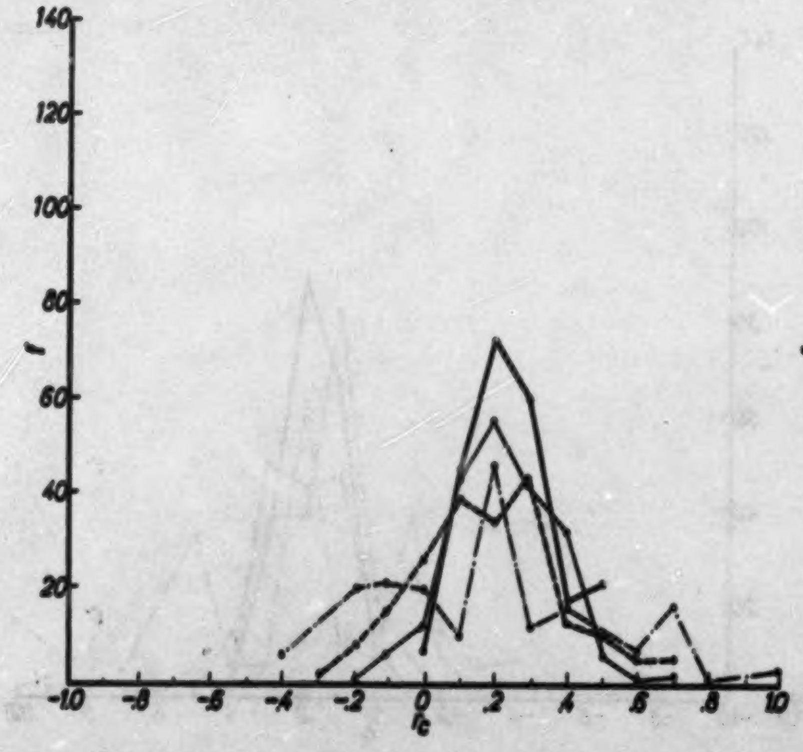


Figure 27g. Item 7, internal criterion, r_c=.208

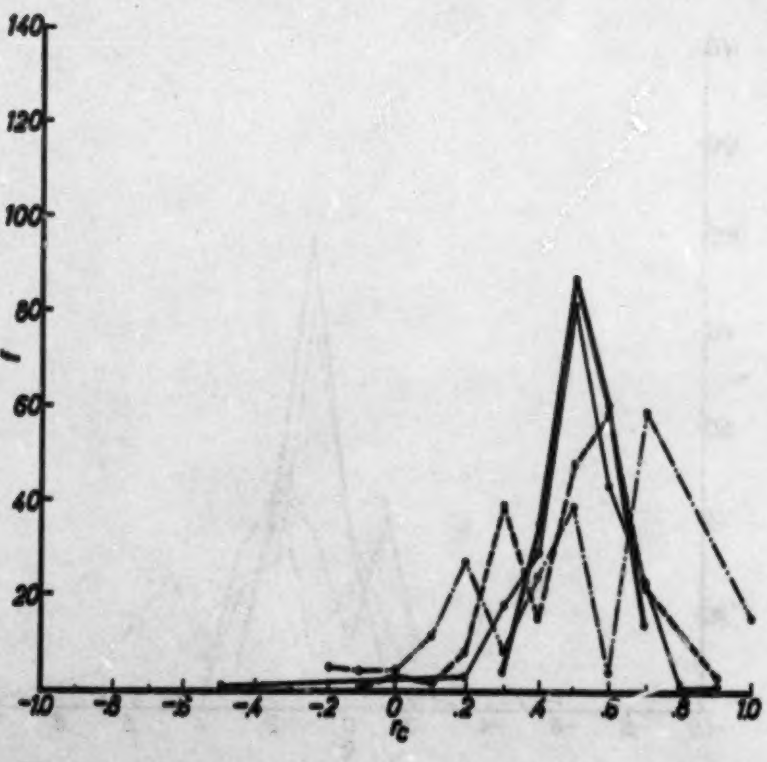


Figure 27h. Item 8, internal criterion, r_c=.532

Figure 27 continued.

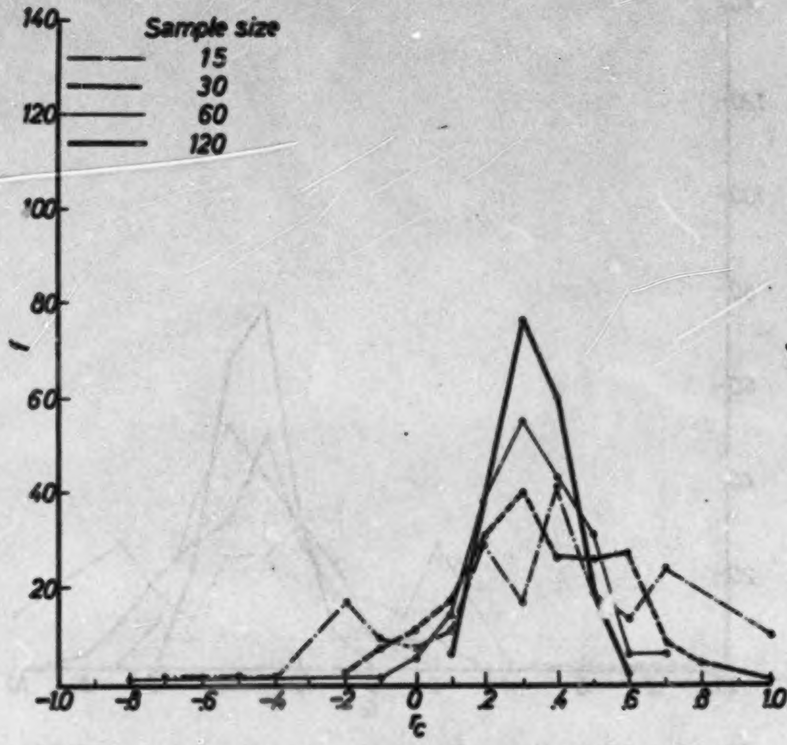


Figure 27i. Item 9, internal criterion, r_C=.347

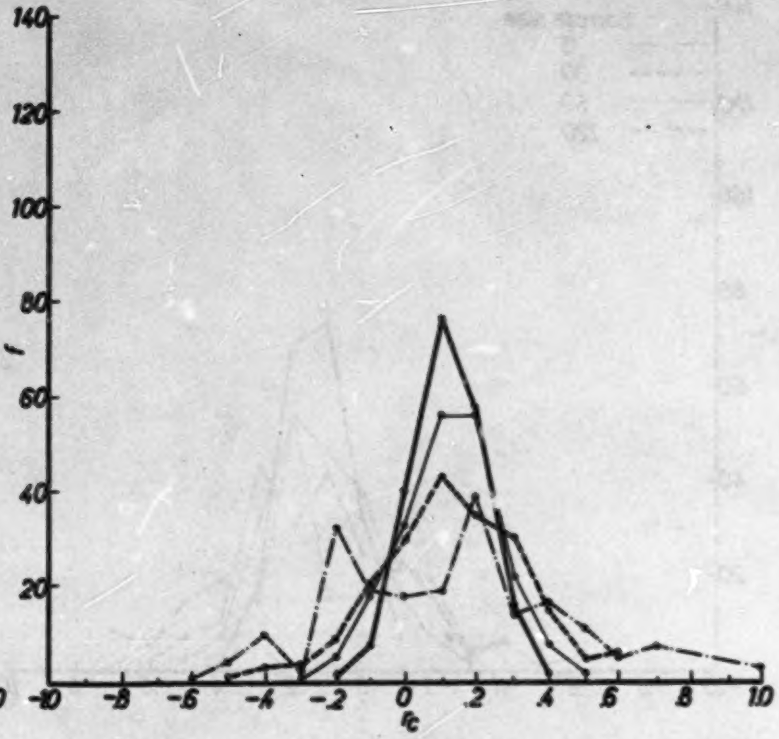


Figure 27j. Item 1, external criterion, r_C=.102

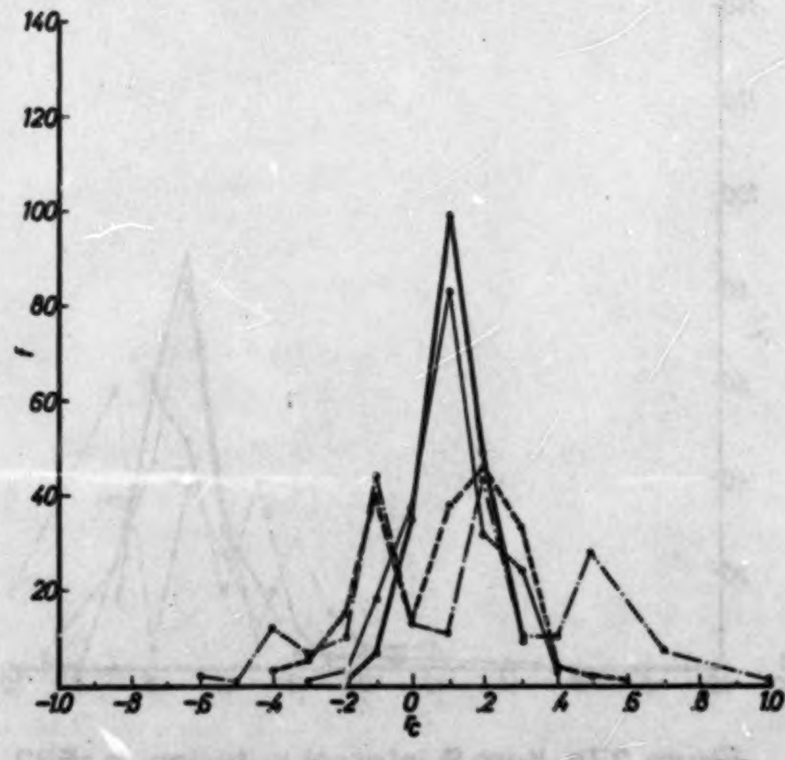


Figure 27k. Item 2, external criterion, r_C=.106

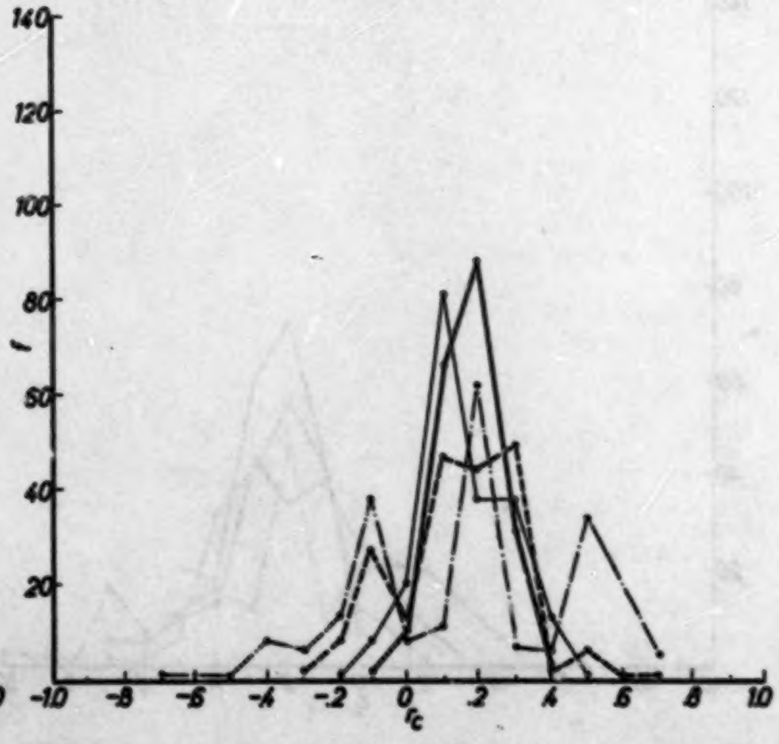


Figure 27l. Item 3, external criterion, r_C=.170

Figure 27 continued.

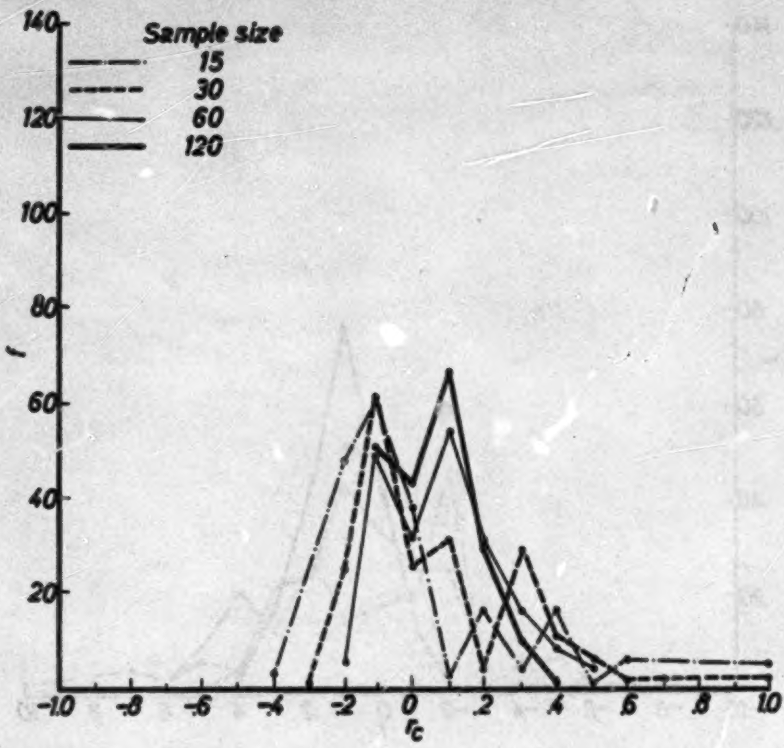


Figure 27m. Item 4, external criterion, r_c=.034

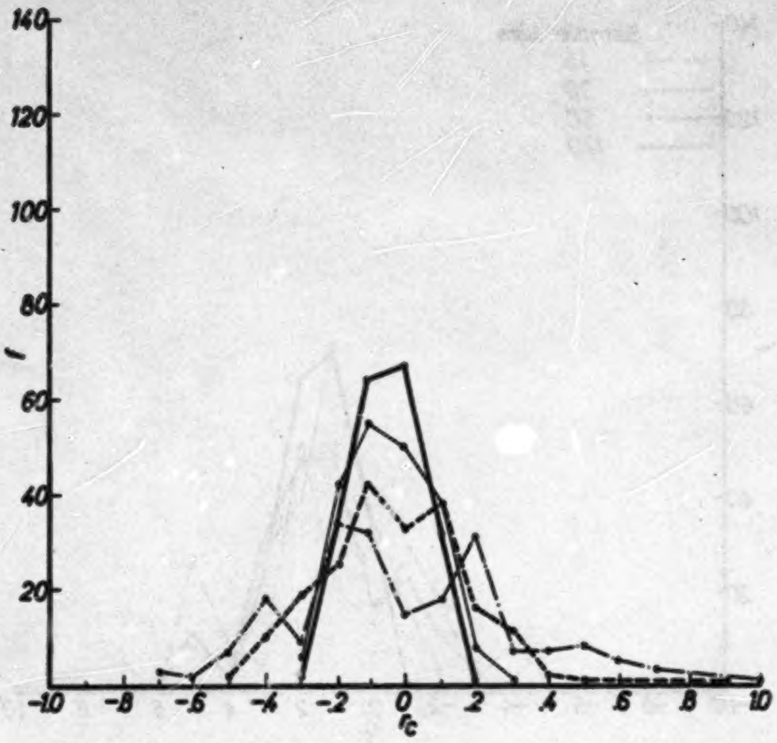


Figure 27n. Item 5, external criterion, r_c=-.050

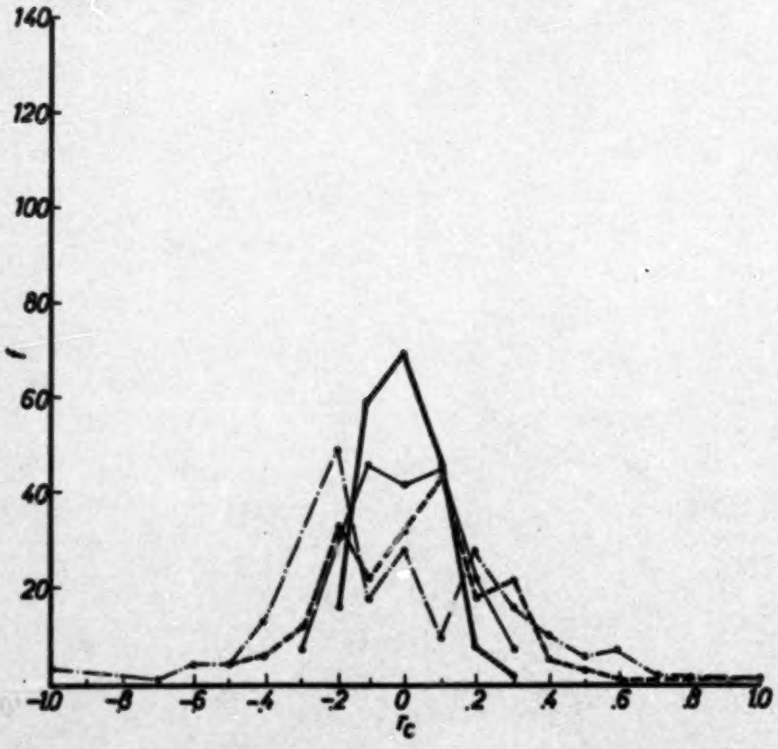


Figure 27o. Item 6, external criterion, r_c=.044

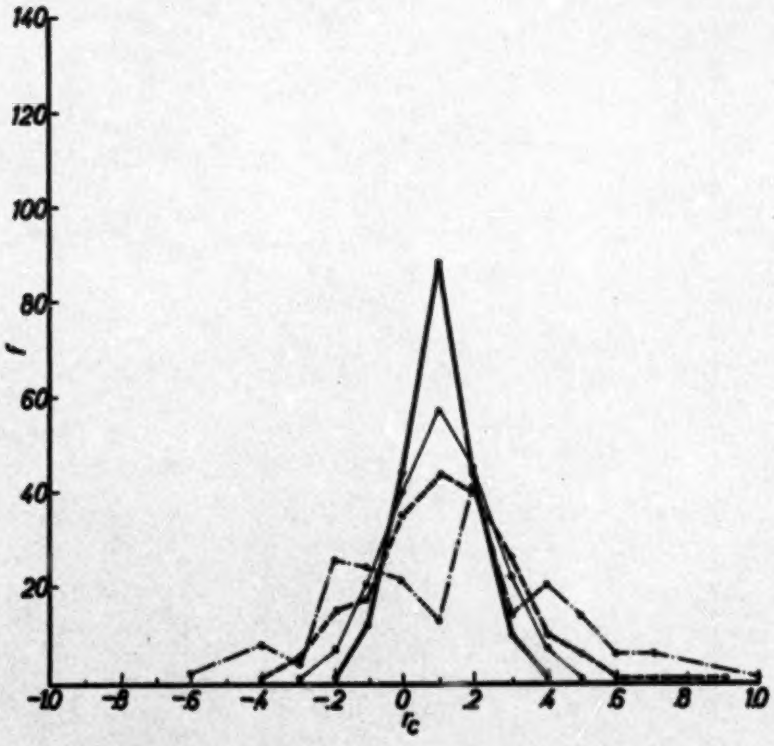


Figure 27p. Item 7, external criterion, r_c=.071

Figure 27 continued.

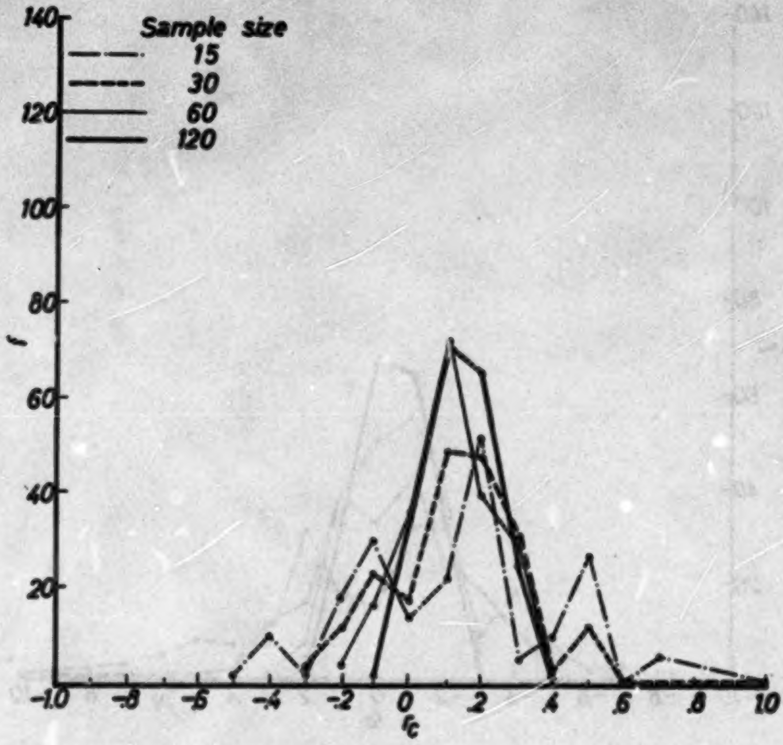


Figure 27q. Item 8, external criterion, r_c=.091

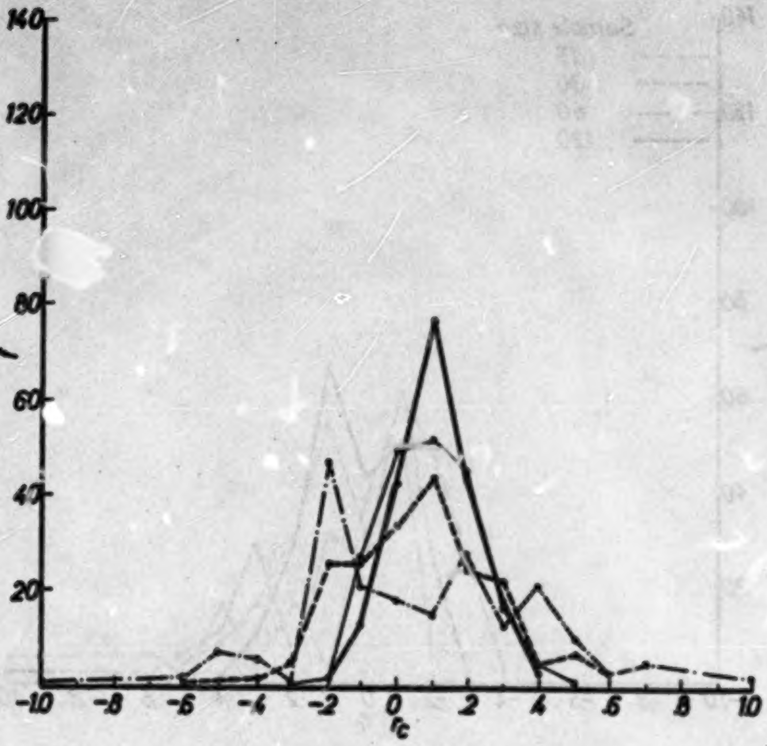


Figure 27r. Item 9, external criterion, r_c=.100

Figure28.Frequency distribution of delta P for the nine items as obtained from random samples of size 15, 30, 60, 120, internal and external criterion.

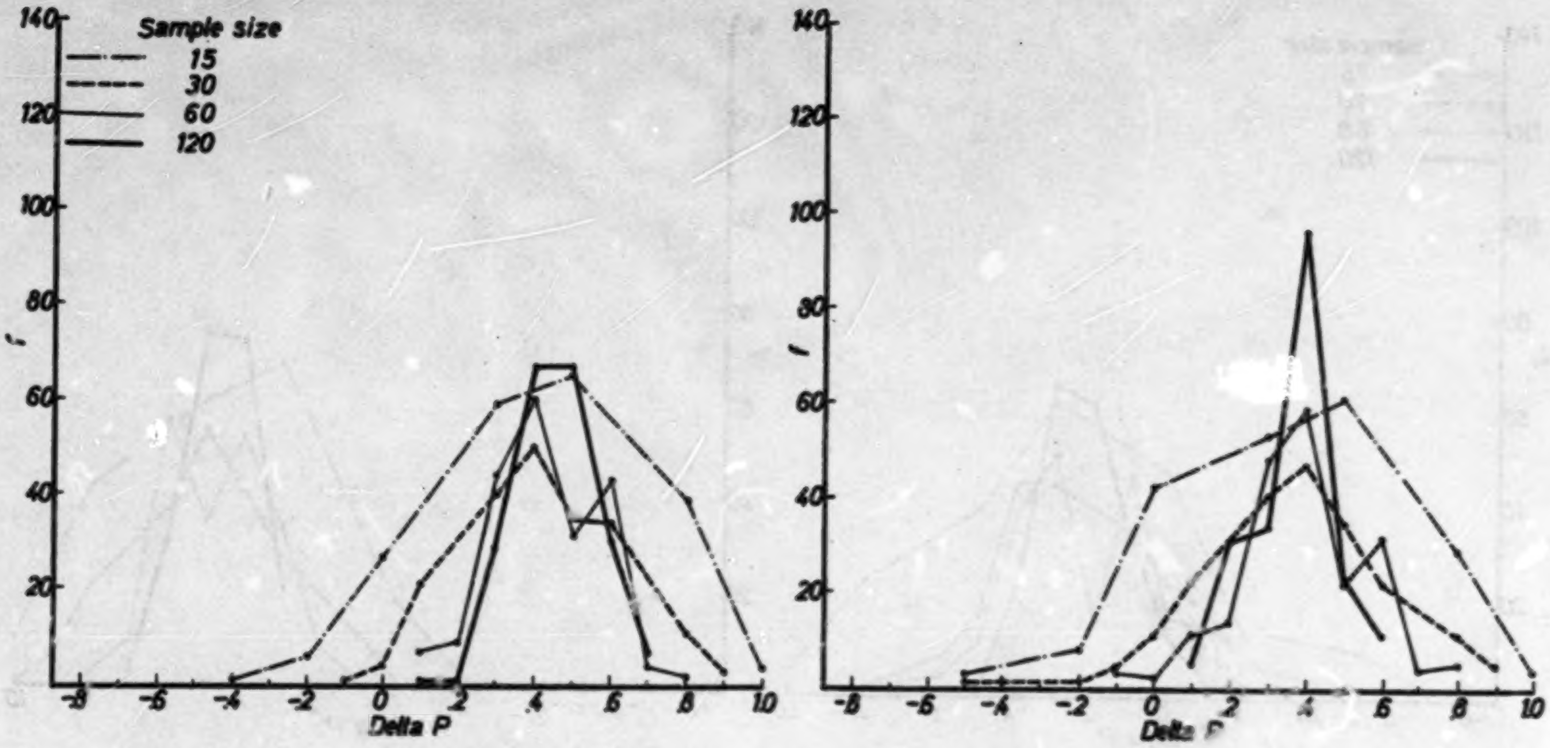


Figure 28a. Item 1 internal criterion Pd=.411

Figure 28b. Item 2 internal criterion Pd=.342

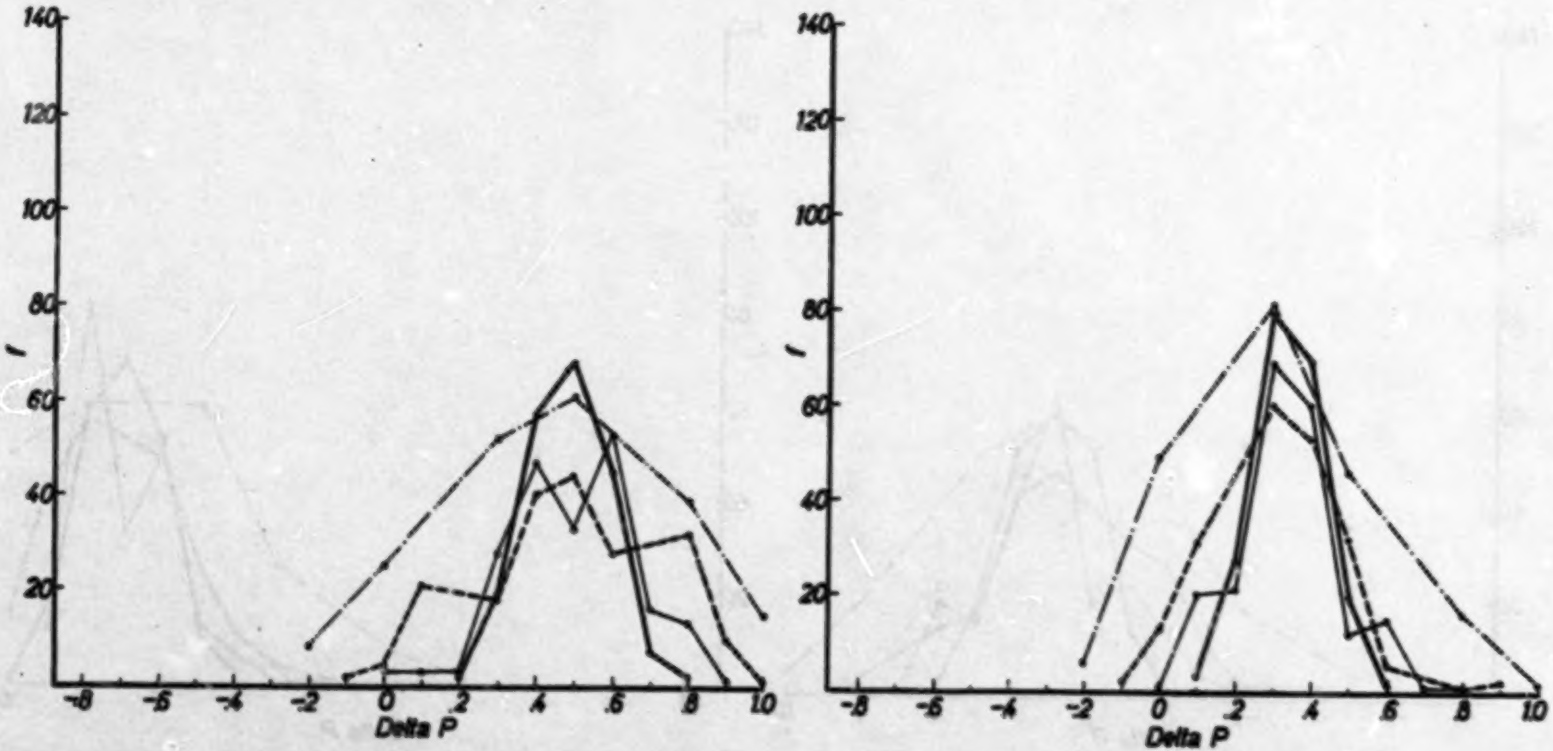


Figure 28c. Item 3 internal criterion Pd=.520

Figure 28d. Item 4 internal criterion Pd=.312

Figure 28 continued.

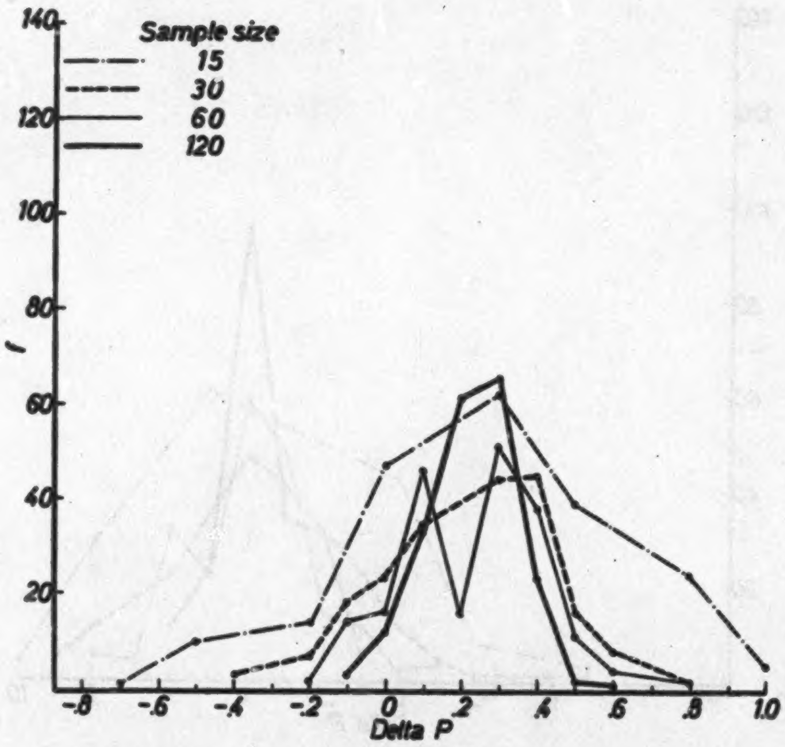


Figure 28e. Item 5, internal criterion, Pd=.198

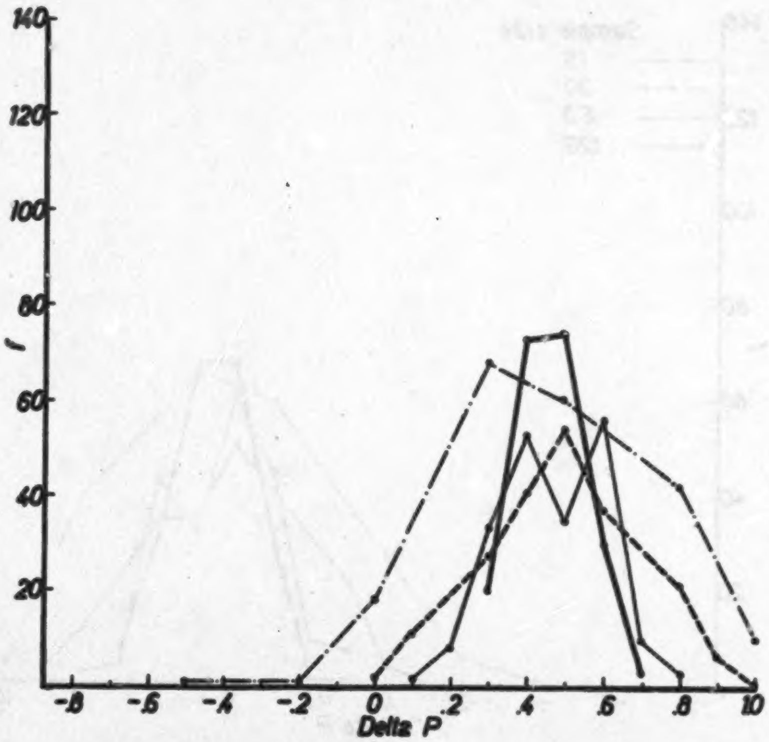


Figure 28f. Item 6, internal criterion, Pd=.451

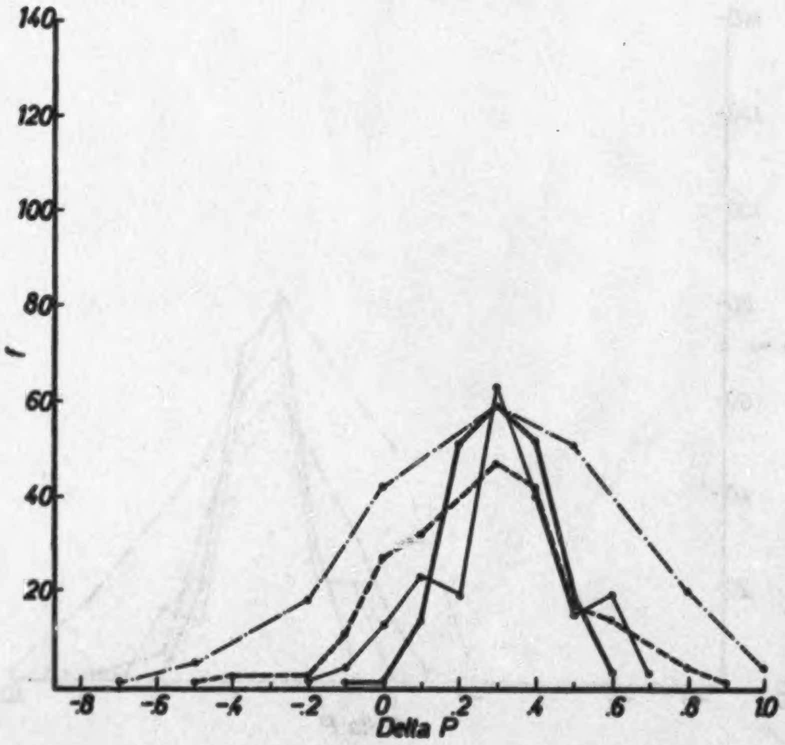


Figure 28g. Item 7, internal criterion, Pd=.282

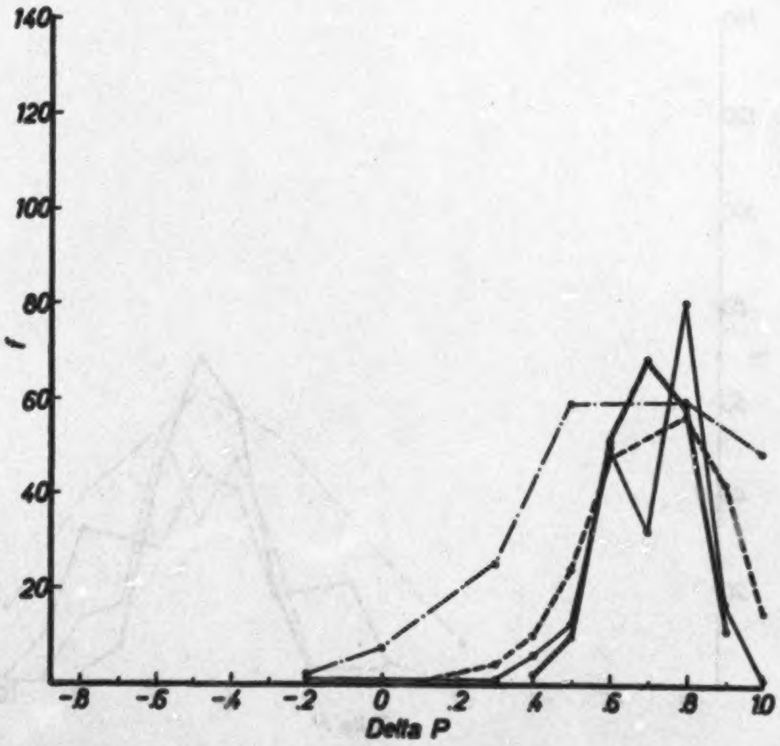


Figure 28h. Item 8, internal criterion, Pd=.683

Figure 28 continued.

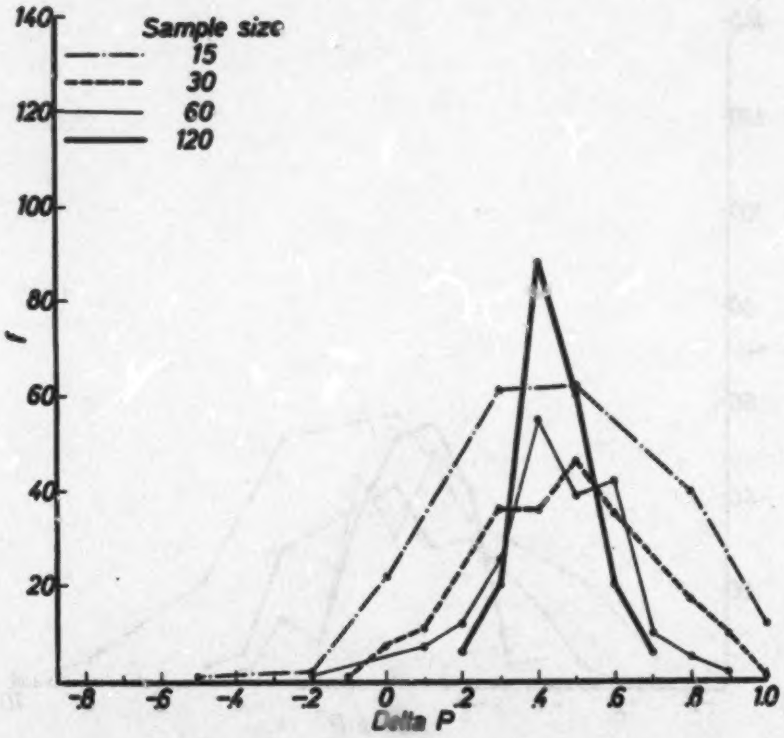


Figure 28i. Item 9, internal criterion, Pd=.450

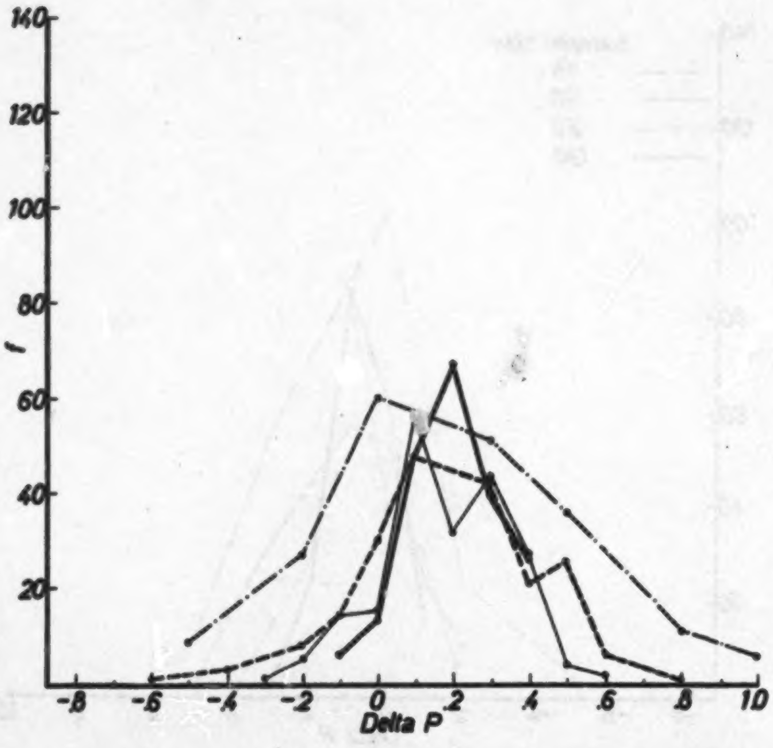


Figure 28j. Item 1, external criterion, Pd=.172

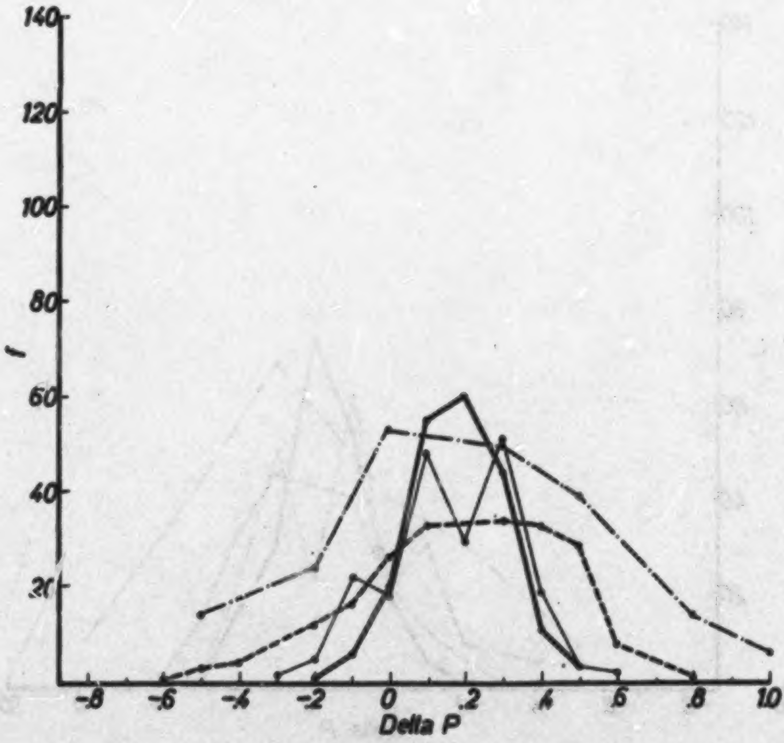


Figure 28k. Item 2, external criterion, Pd=.172

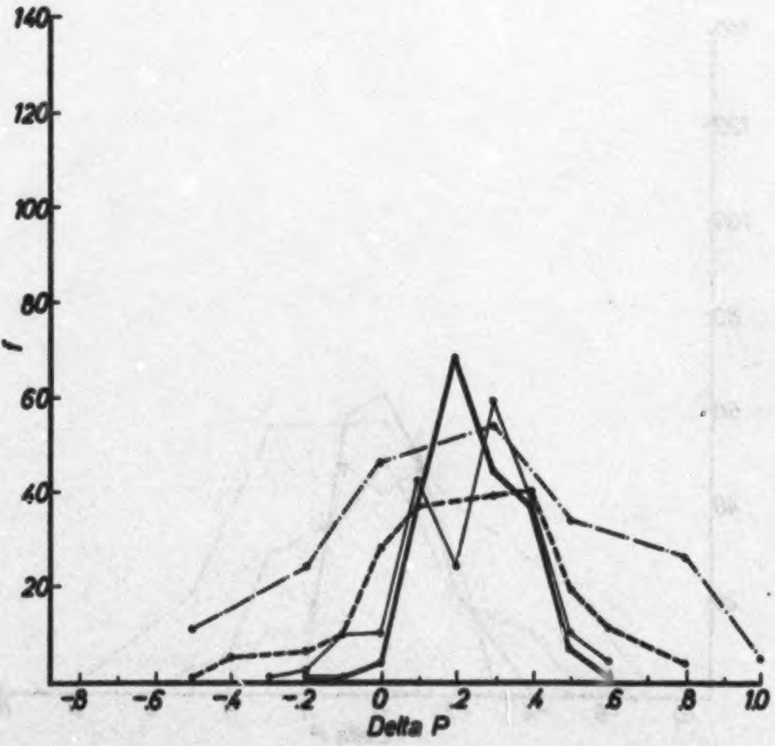


Figure 28l. Item 3, external criterion, Pd=.227

Figure 28 continued.

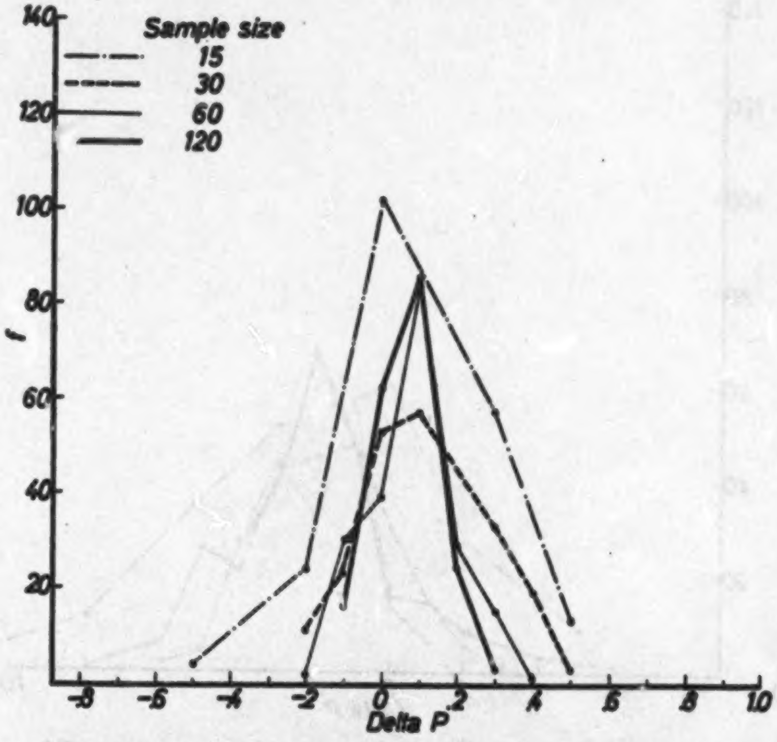


Figure 28m. Item 4, external criterion, Pd=.074

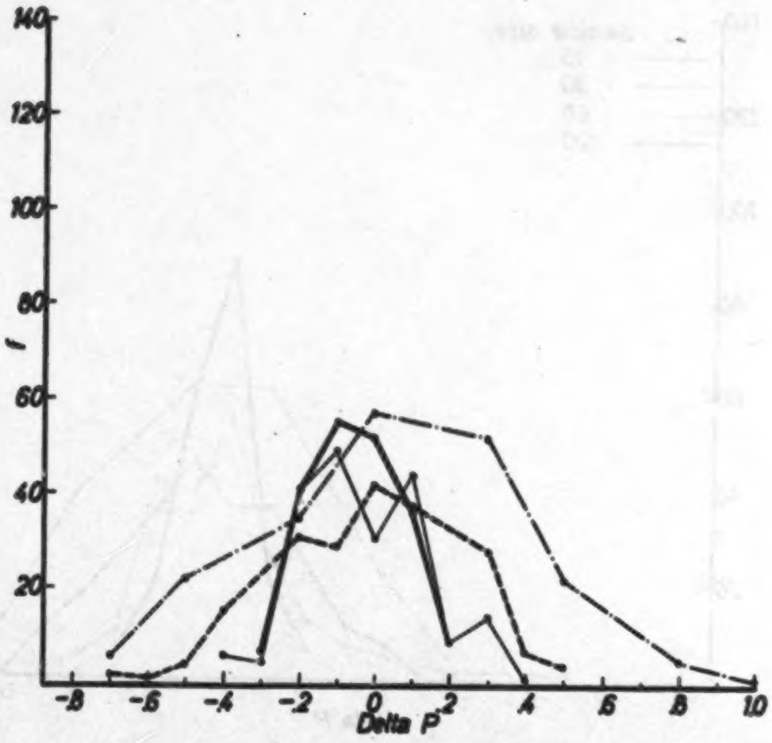


Figure 28n. Item 5, external criterion, Pd=-.020

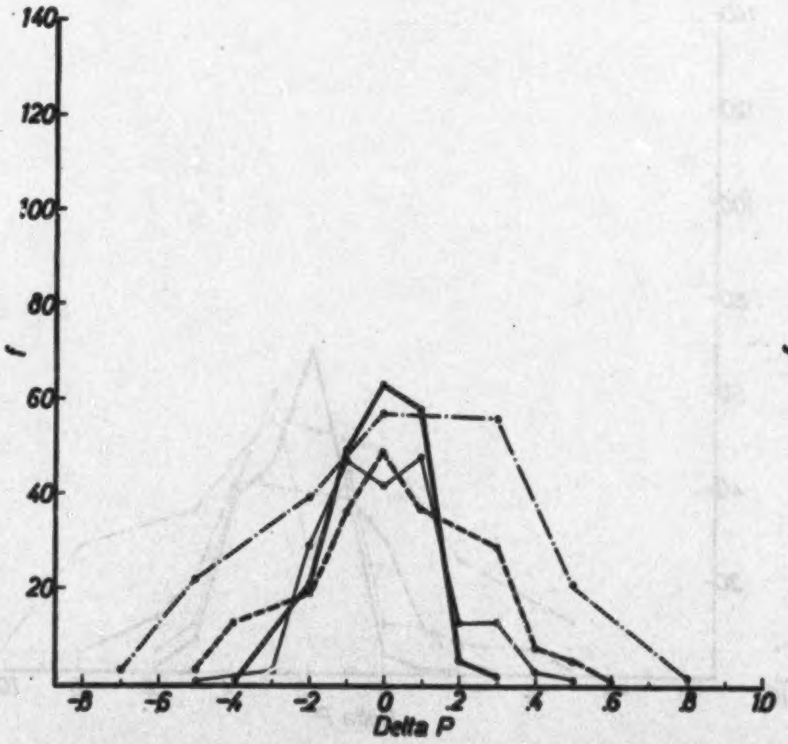


Figure 28o. Item 6, external criterion, Pd=.015

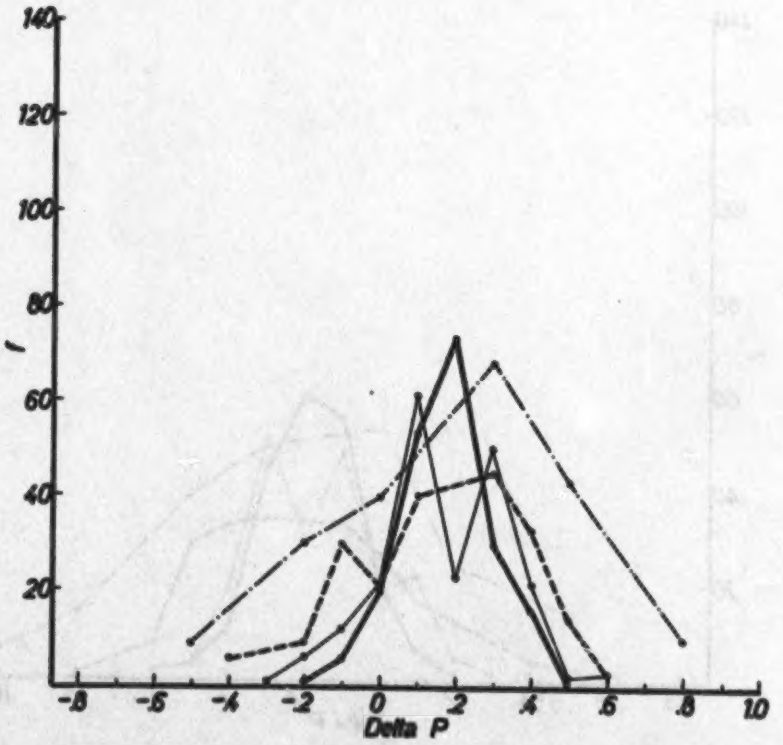


Figure 28p. Item 7, external criterion, Pd=.173

Figure 28 continued.

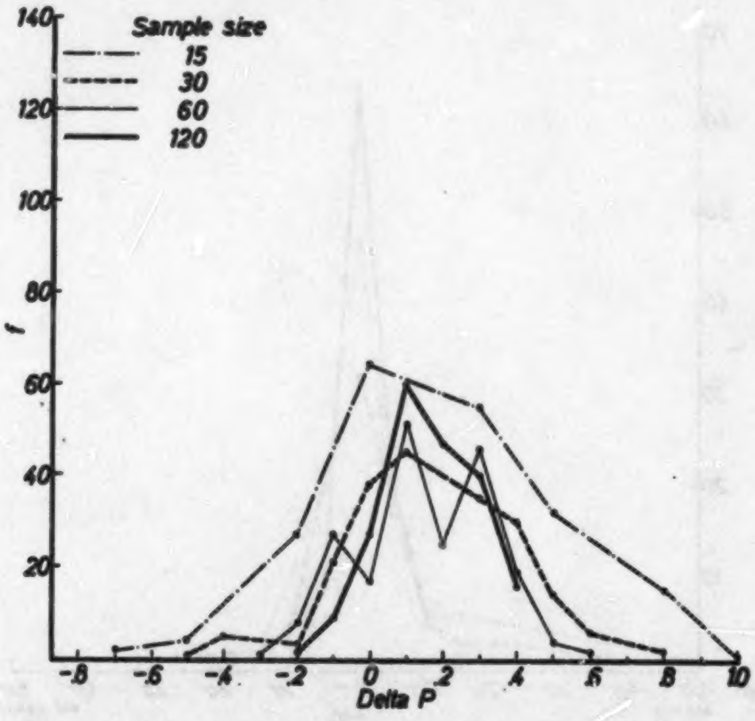


Figure 28q. Item 8, external criterion, Pd=.120

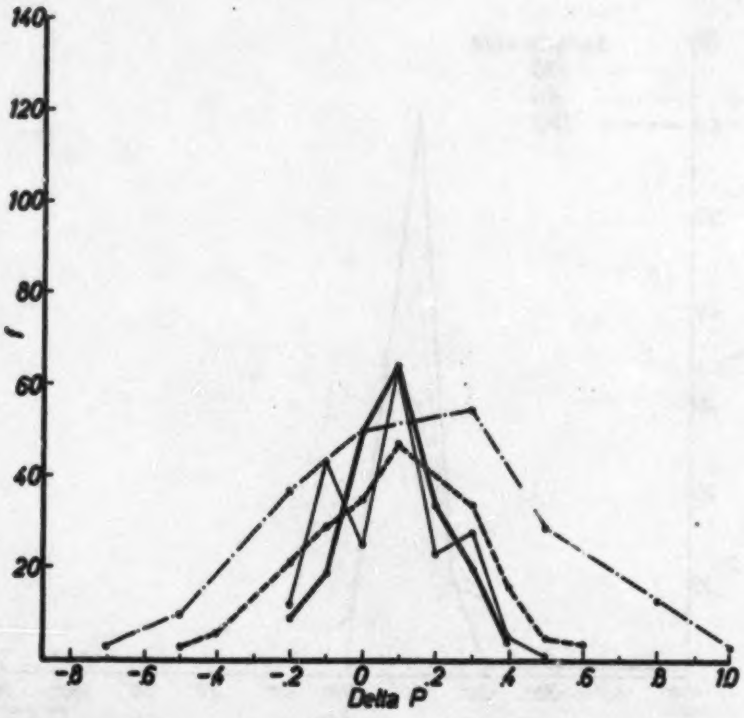


Figure 28r. Item 9, external criterion, Pd=.090

Figure 29. Frequency distribution of the maximum likelihood estimate of the item parameter X_{50} for the nine items as obtained from random samples of size 30, 60, 120, internal and external criterion.

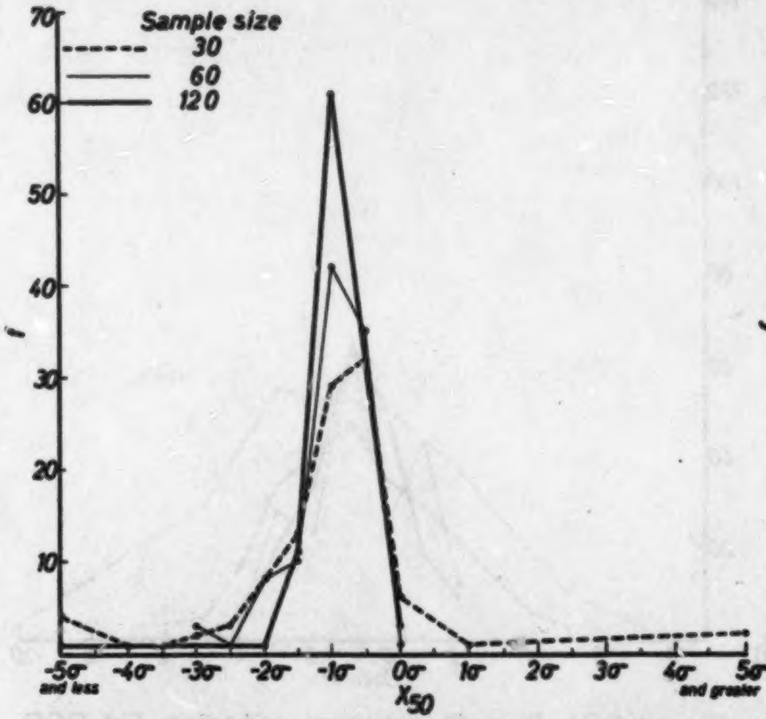


Figure 29a. Item 1, internal criterion, X_{50}=-1068

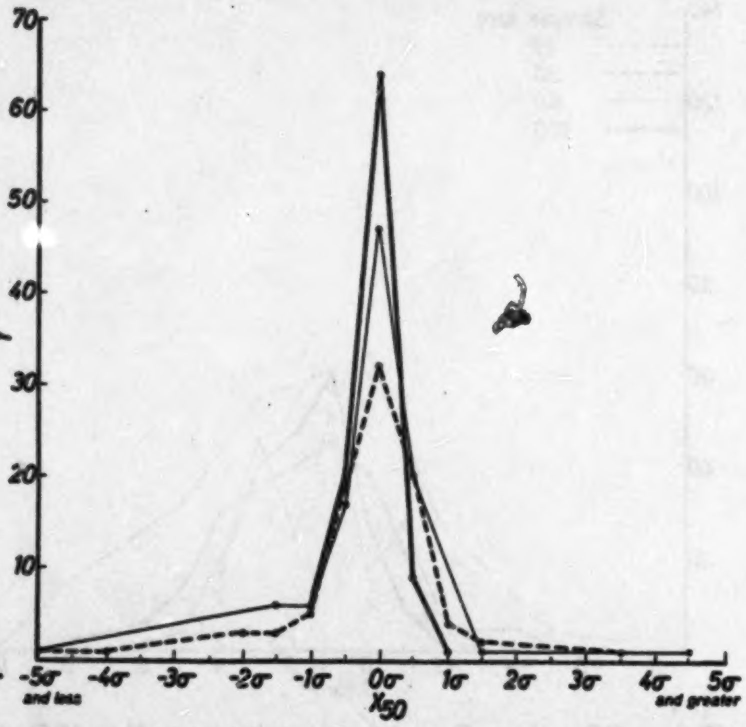


Figure 29b. Item 2 internal criterion X_{50}=-.050

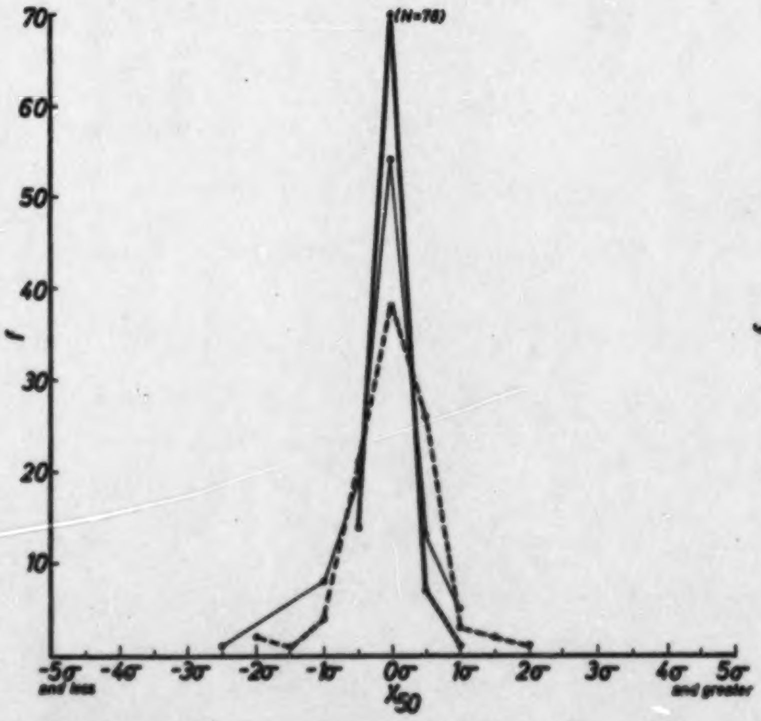


Figure 29c. Item 3 internal criterion X_{50}=.011

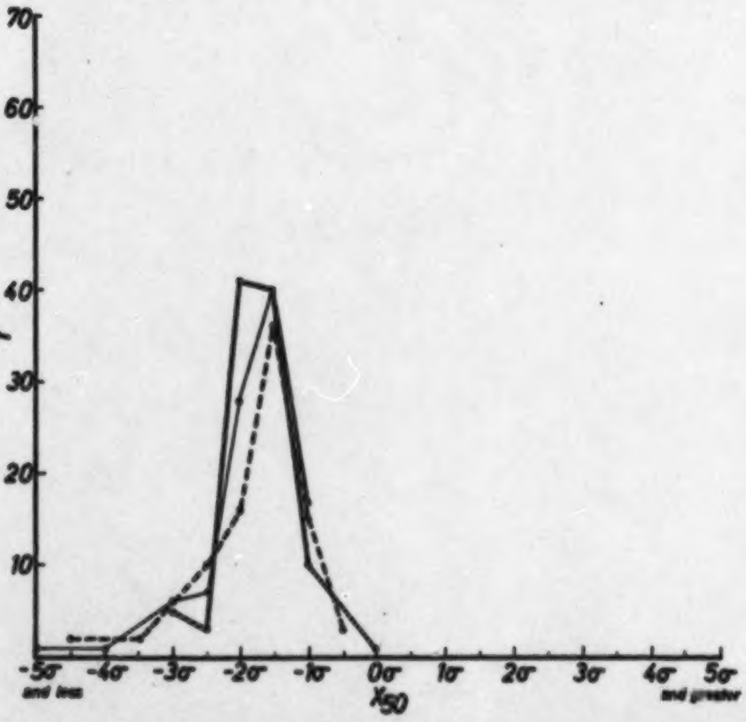


Figure 29d. Item 4 internal criterion X_{50}=-1753

Figure 29 continued.

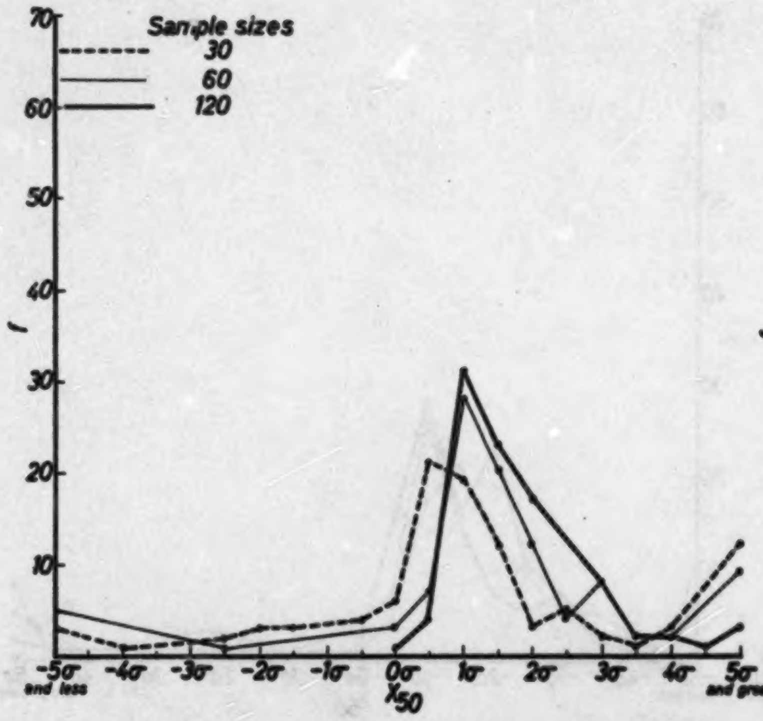


Figure 29e. Item 5, internal criterion, X_{50}=1.868

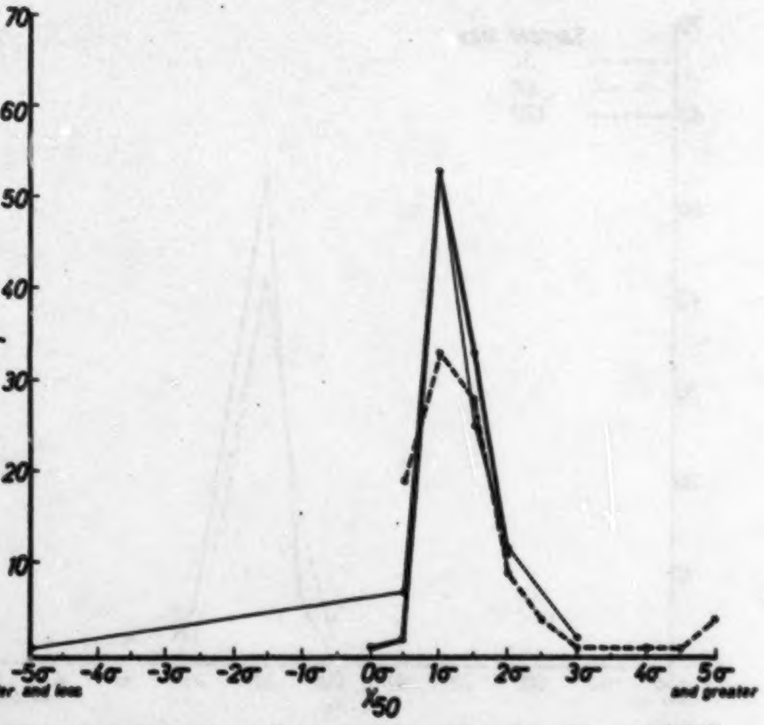


Figure 29f. Item 6, internal criterion, X_{50}=1.290

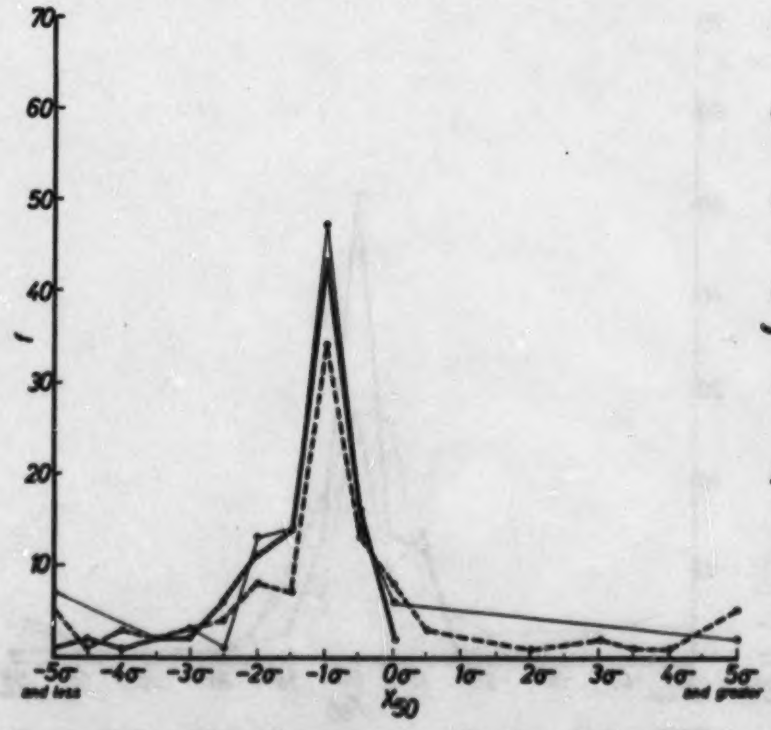


Figure 29g. Item 7, internal criterion, X_{50}=1.204

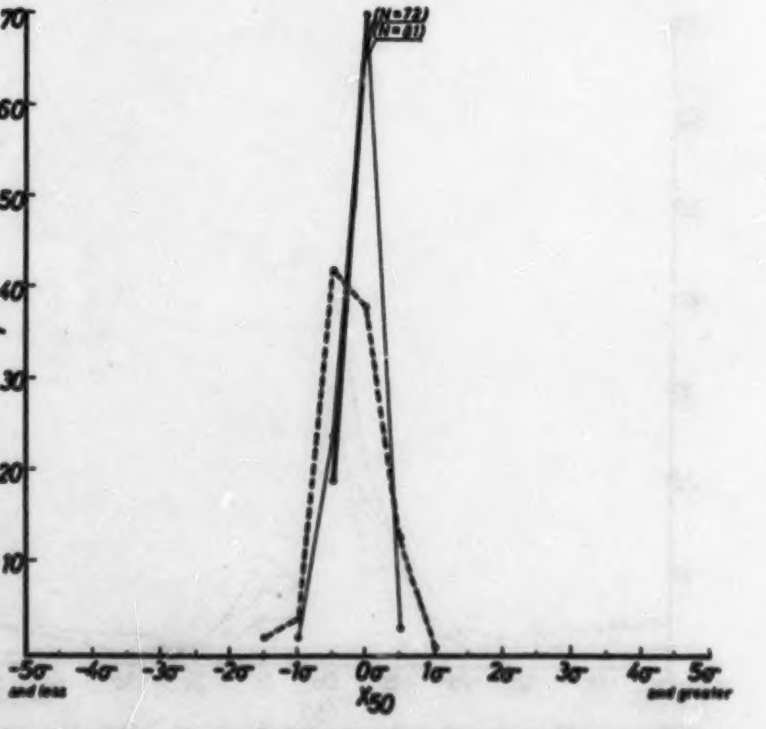


Figure 29h. Item 8, internal criterion, X_{50}=-.136

Figure 29 continued.

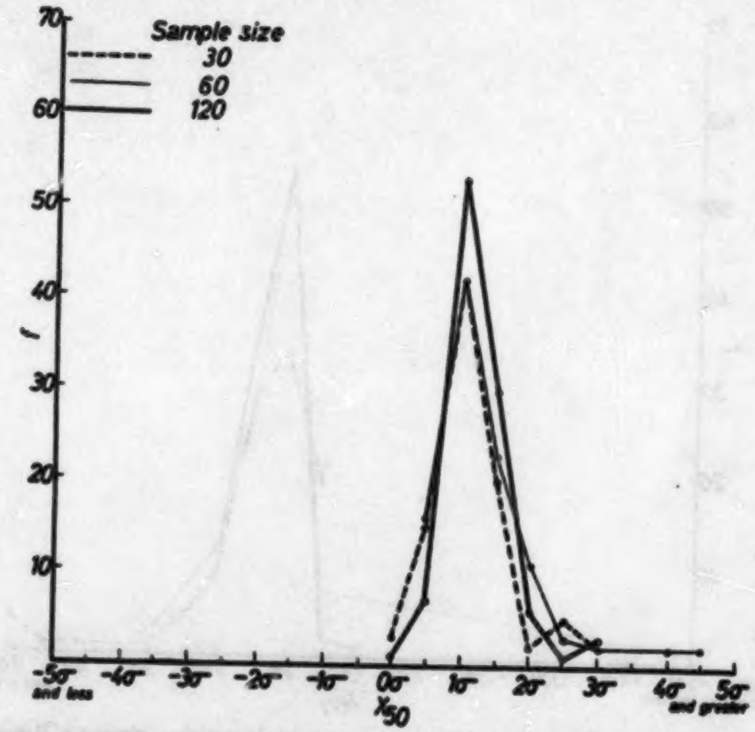


Figure 29i. Item 9, internal criterion, X_{50}=1.169

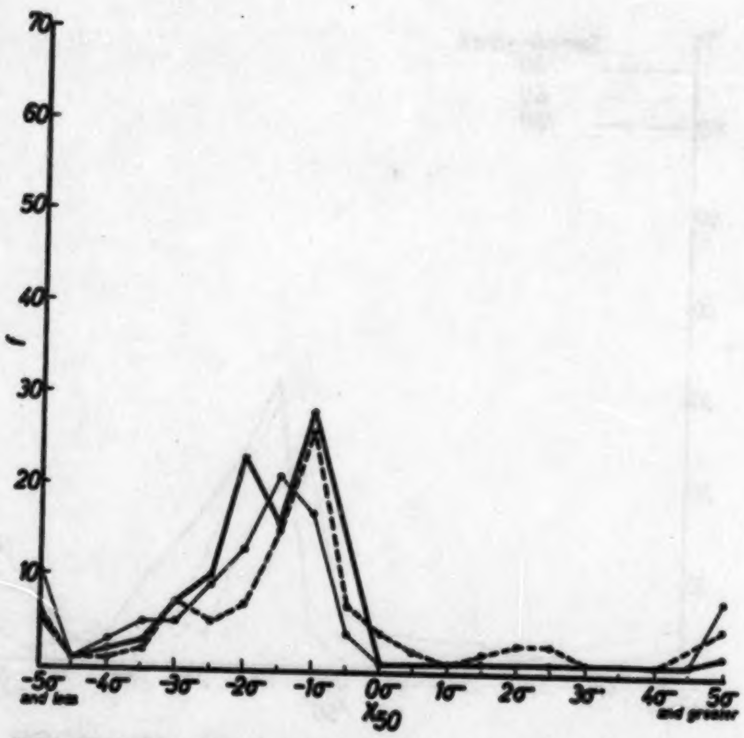


Figure 29j. Item 1, external criterion, X_{50}=-2.506

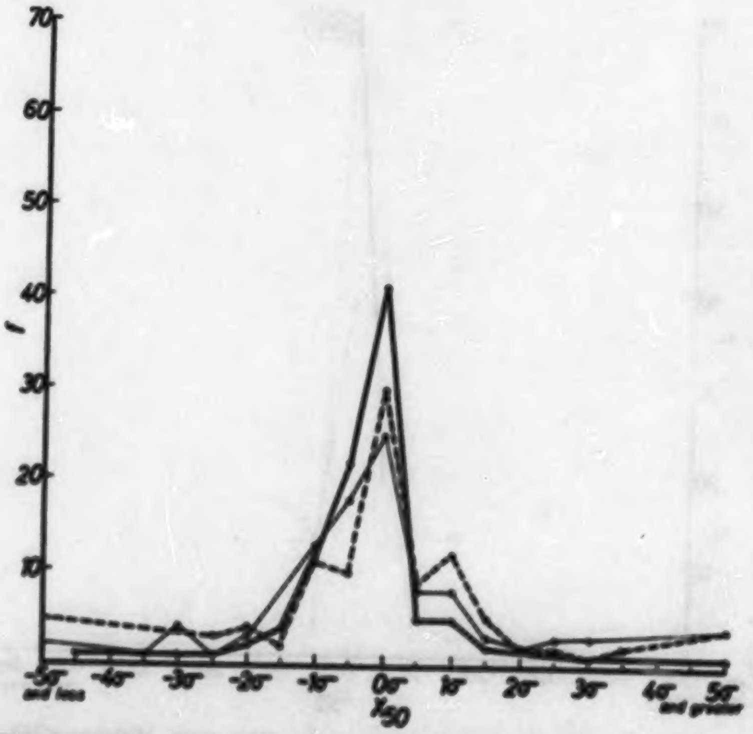


Figure 29k. Item 2, external criterion, X_{50}=-.077

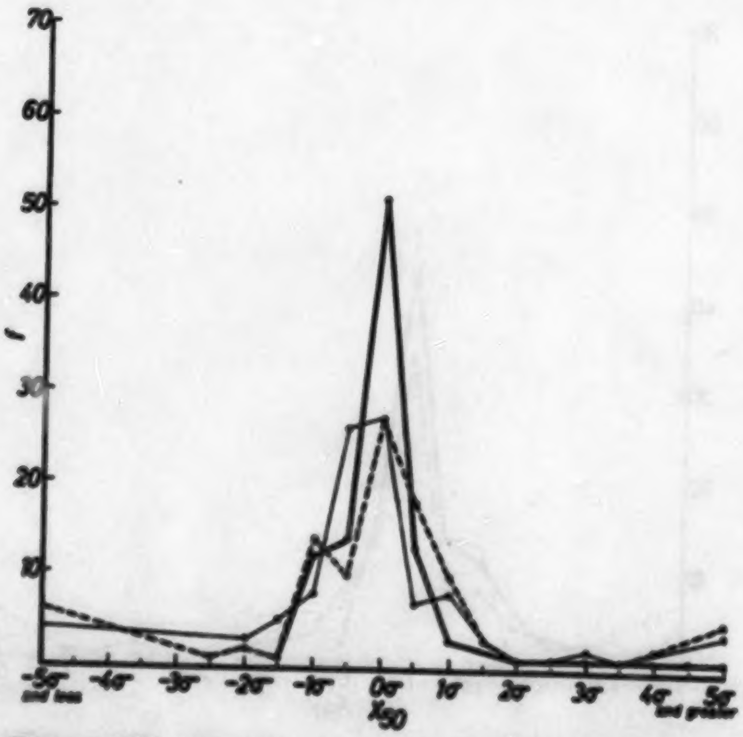


Figure 29l. Item 3, external criterion, X_{50}=.032

Figure 29 continued.

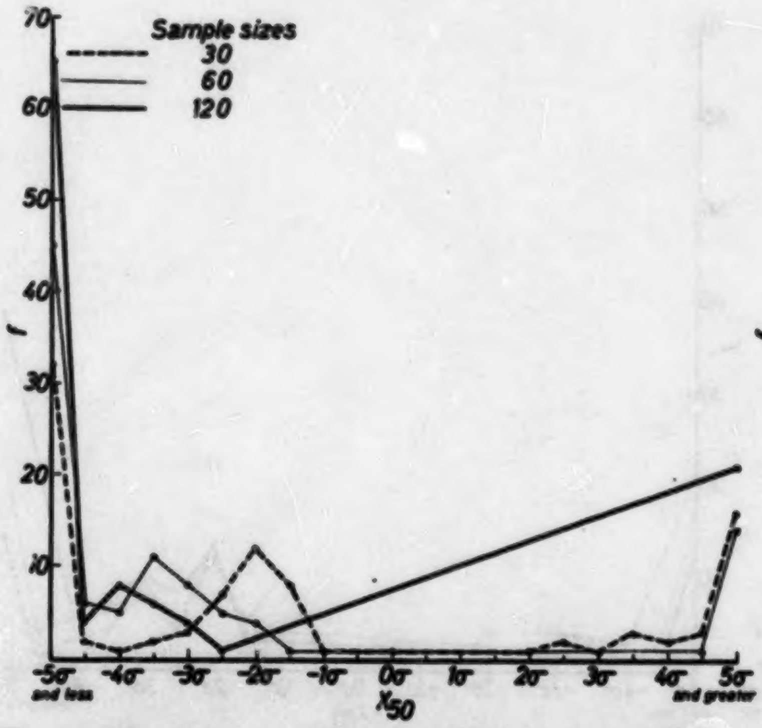


Figure 29m. Item 4, external criterion, X_{50}=-6.119

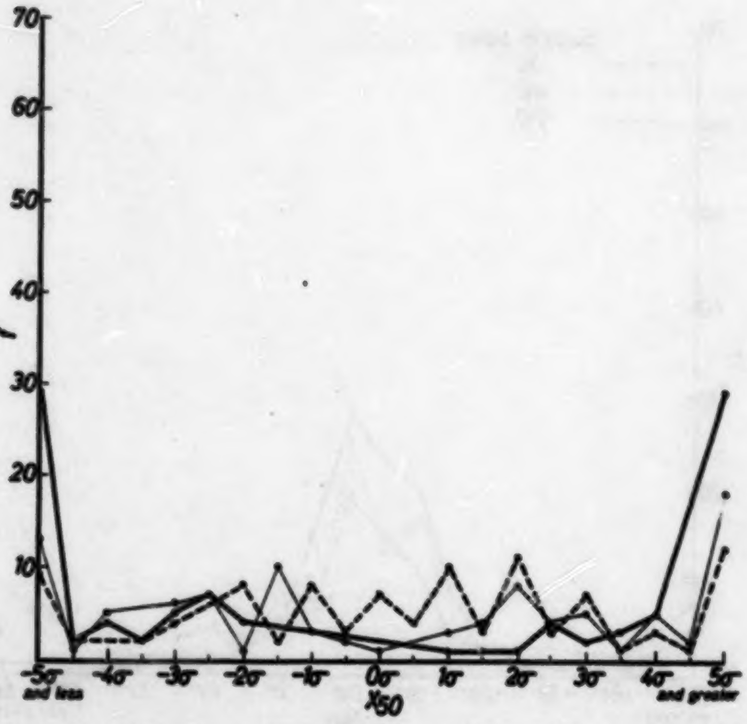


Figure 29n. Item 5, external criterion, X_{50}=-34.555

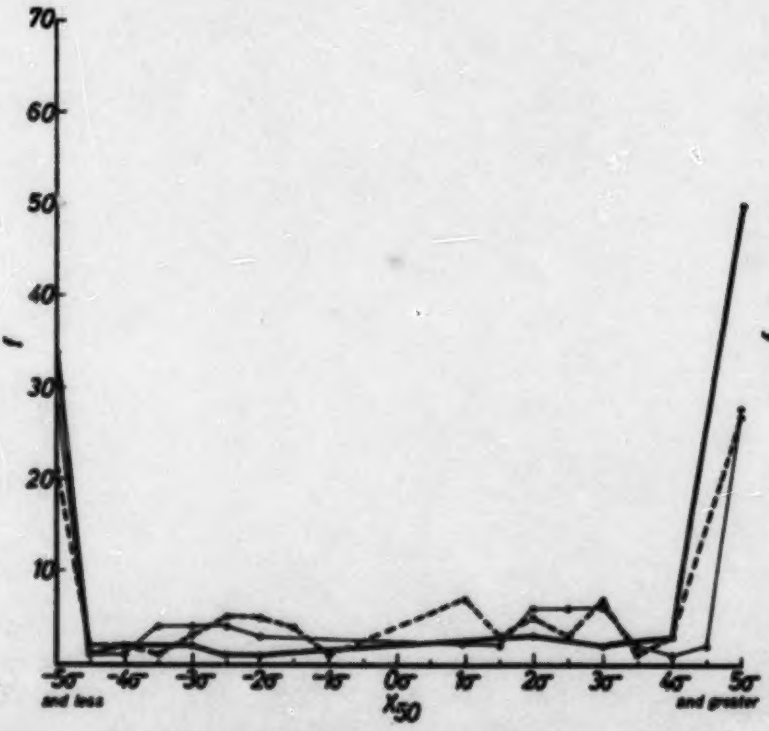


Figure 29o. Item 6, external criterion, X_{50}=11.050

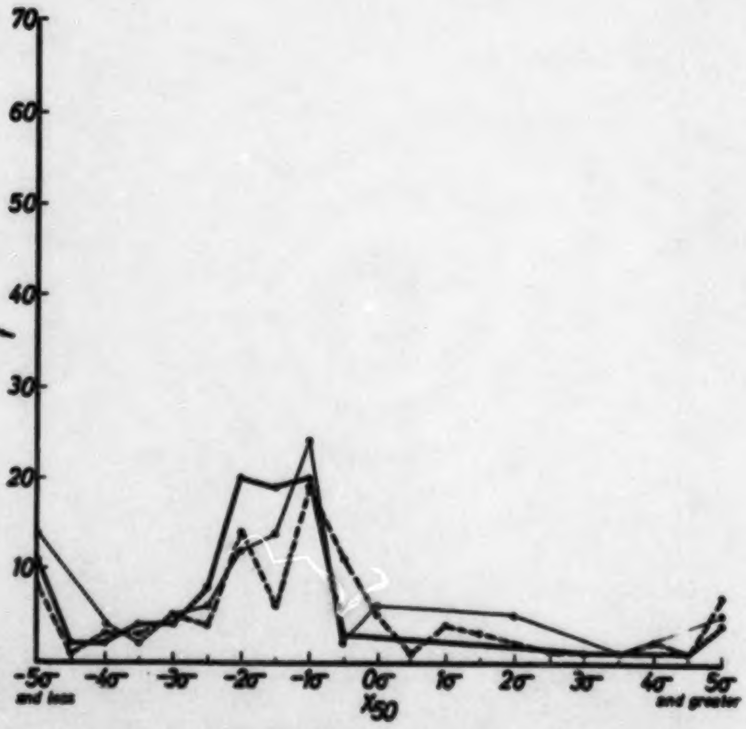


Figure 29p. Item 7, external criterion, X_{50}=-2.170

Figure 29 continued.

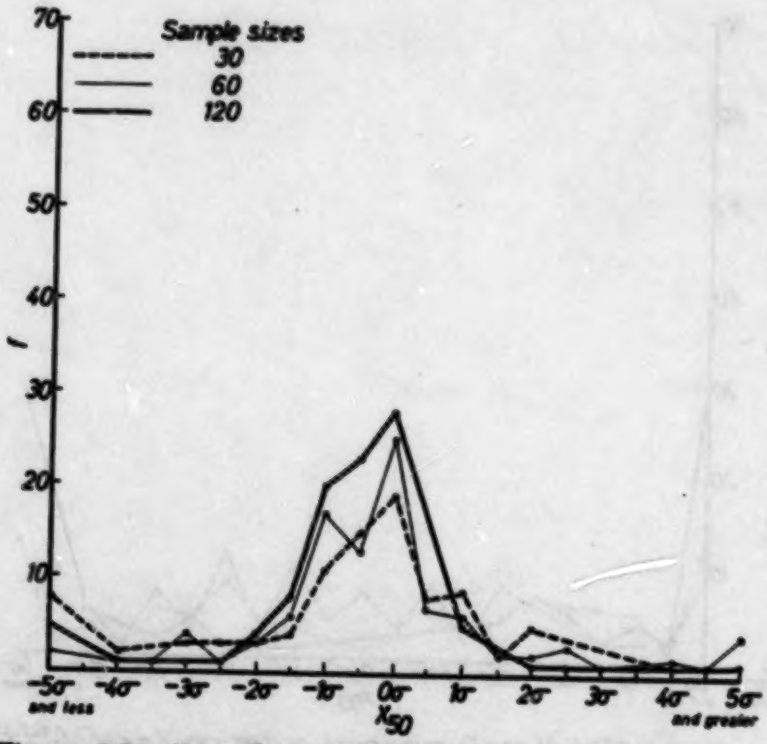


Figure 29q. Item 8, external criterion, $X_{50}=-.603$

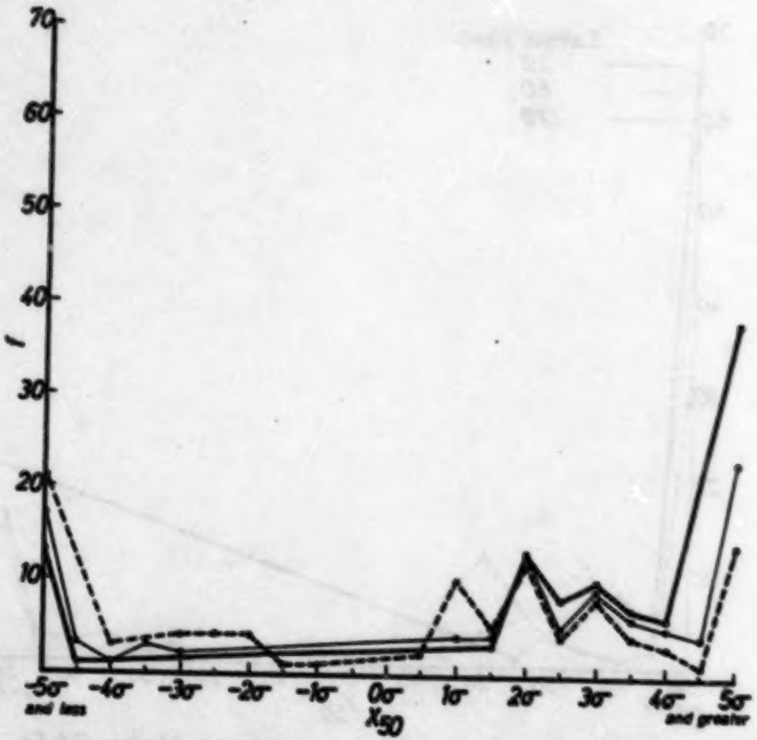


Figure 29r. Item 9, external criterion, $X_{50}=4.266$

Figure 30. Frequency distribution of the maximum likelihood estimate of the item parameter β for the nine items as odtained from random samples of size 30, 60, 120, internal and external criterion.

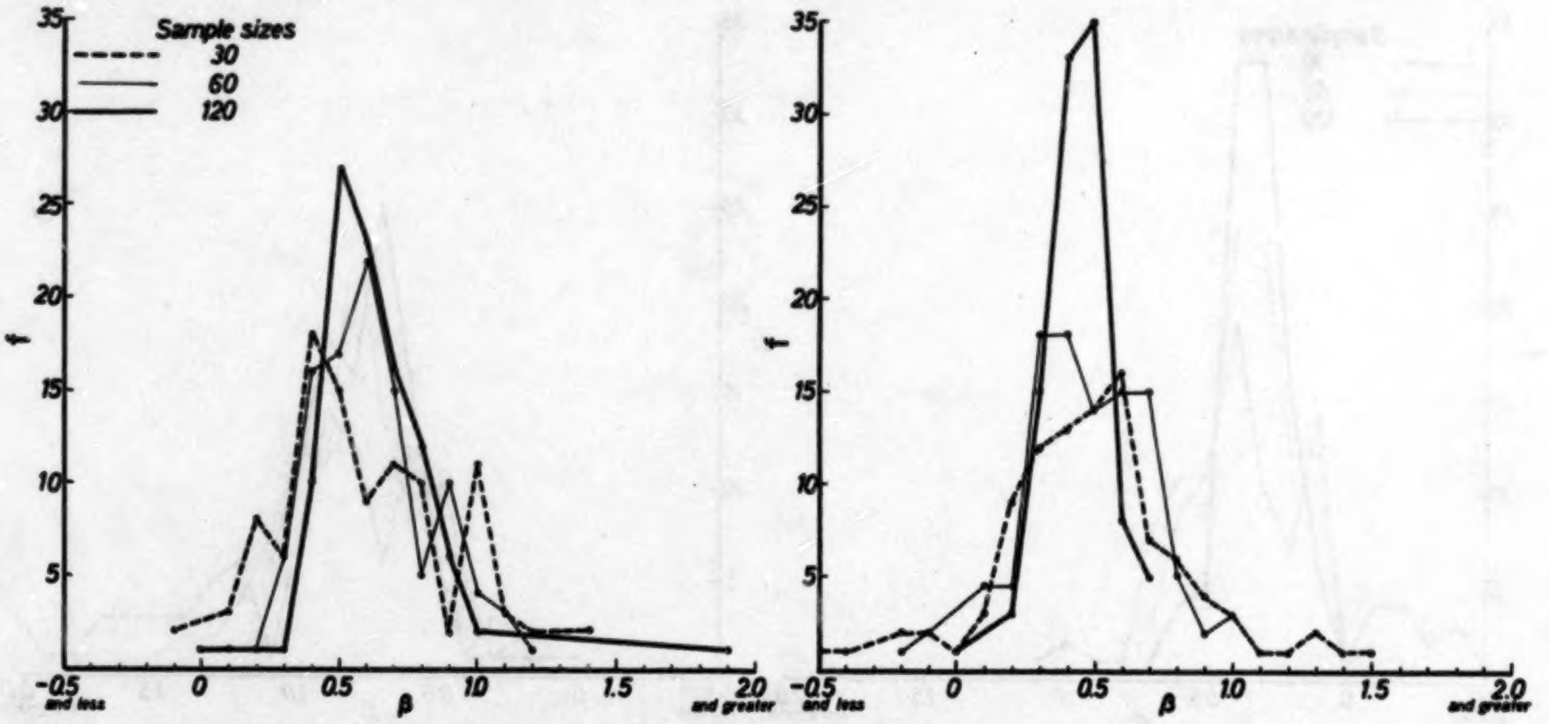


Figure 30a. Item 1, internal criterion, $\beta = .532$ Figure 30b. Item 2, internal criterion, $\beta = .358$

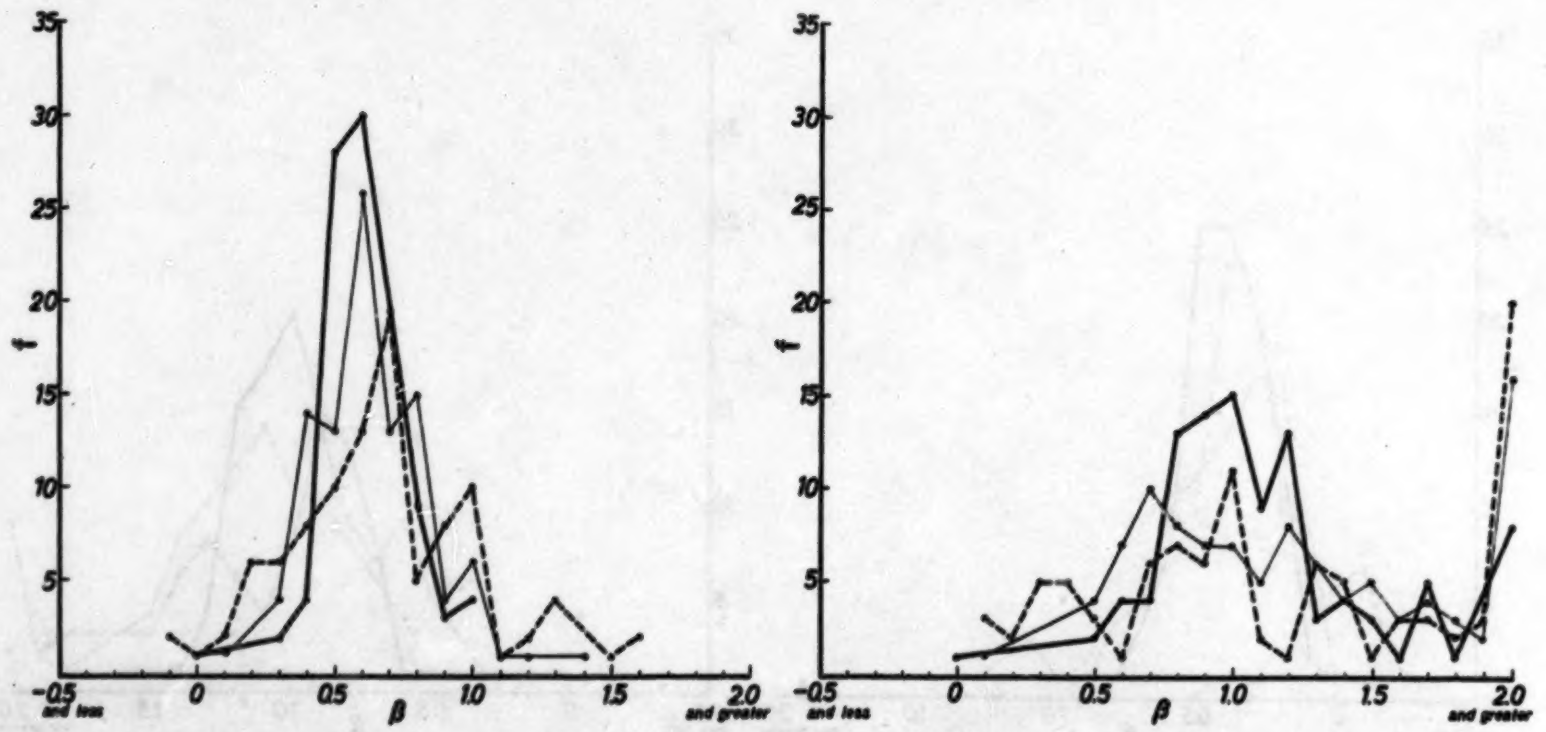


Figure 30c. Item 3, internal criterion, $\beta = .610$ Figure 30d. Item 4, internal criterion, $\beta = .954$

Figure 30 continued.

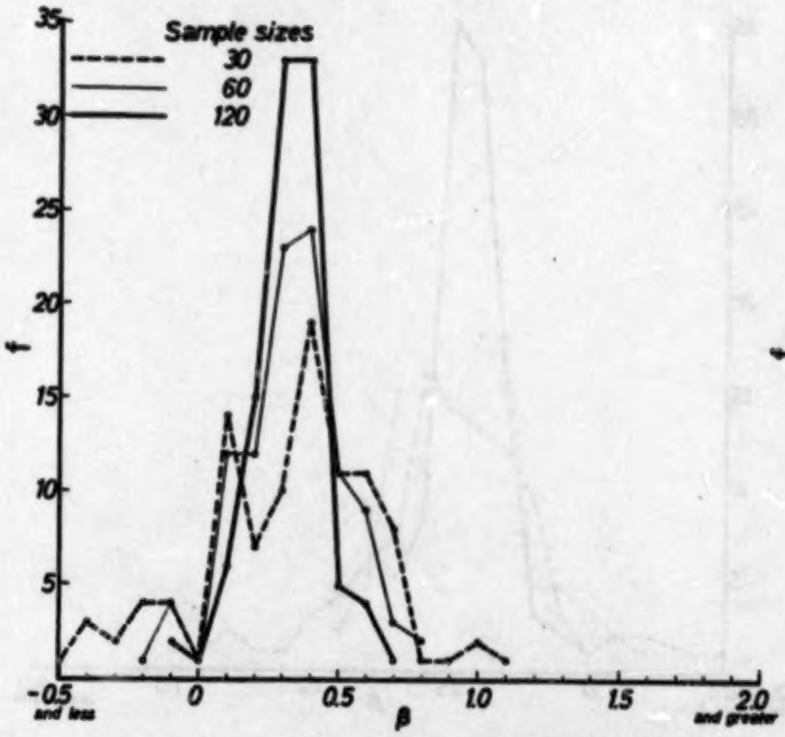


Figure 30e. Item 5, internal criterion, β =.239

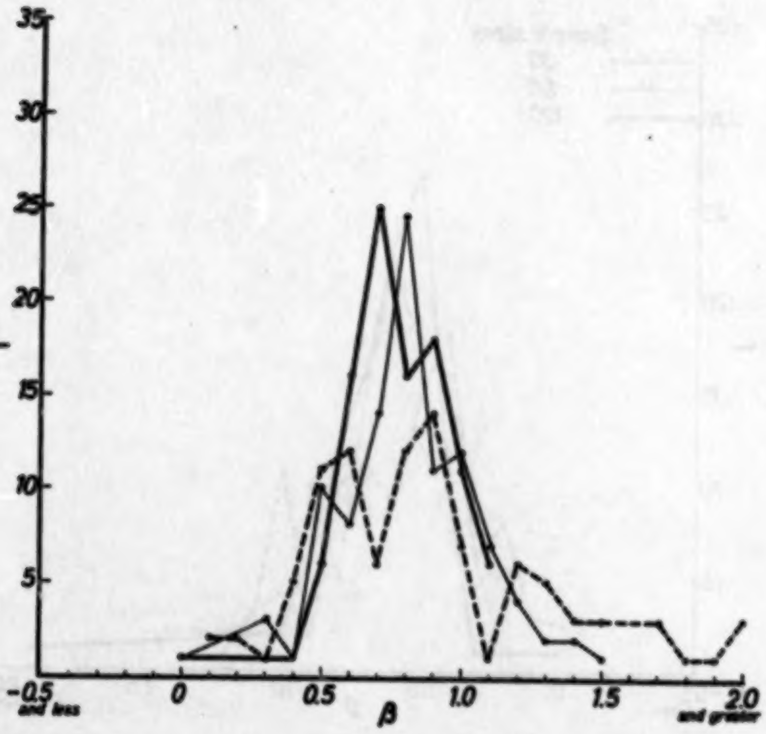


Figure 30f. Item 6, internal criterion, β =.714

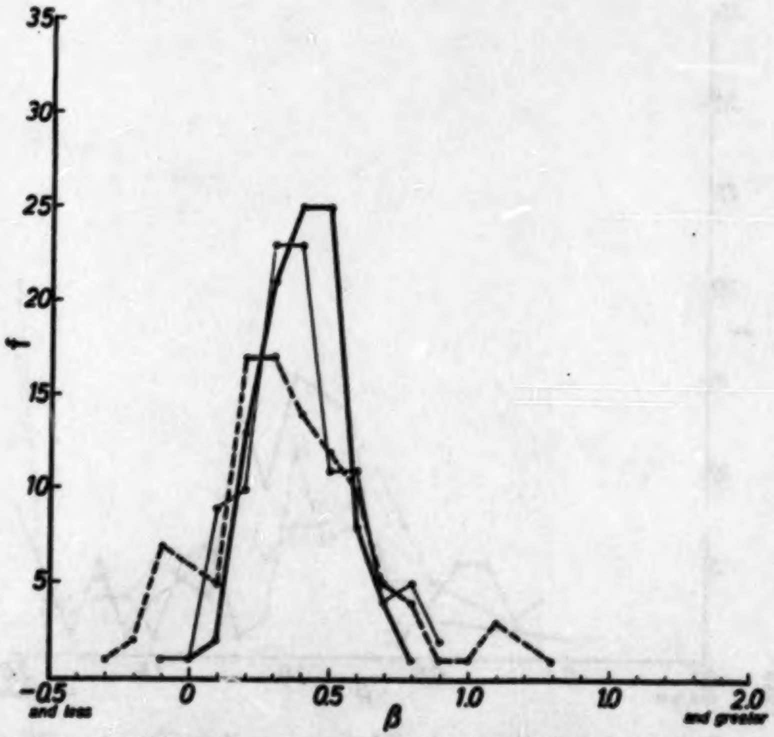


Figure 30g. Item 7, internal criterion, β =.360

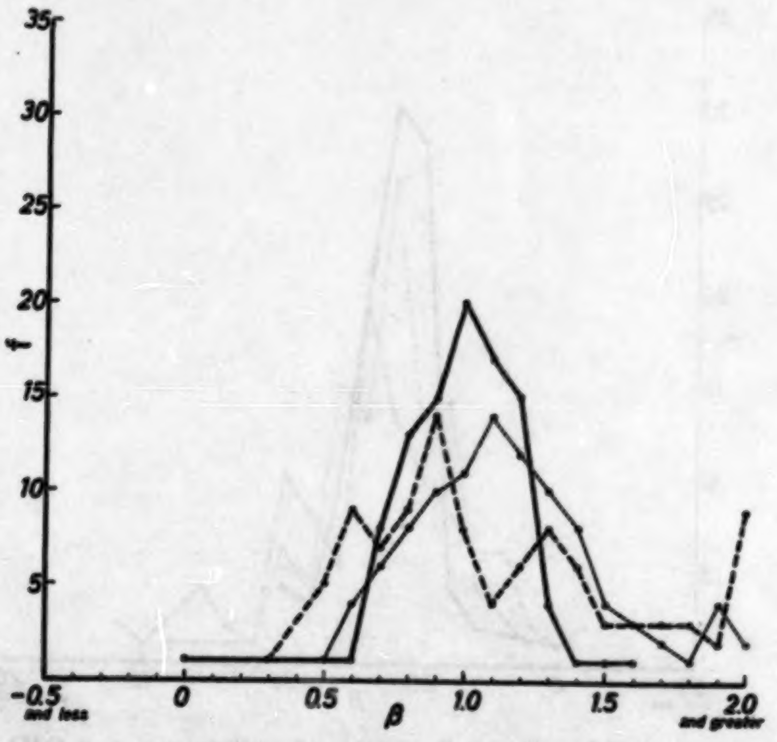


Figure 30h. Item 8, internal criterion, β =.929

Figure 30 continued.

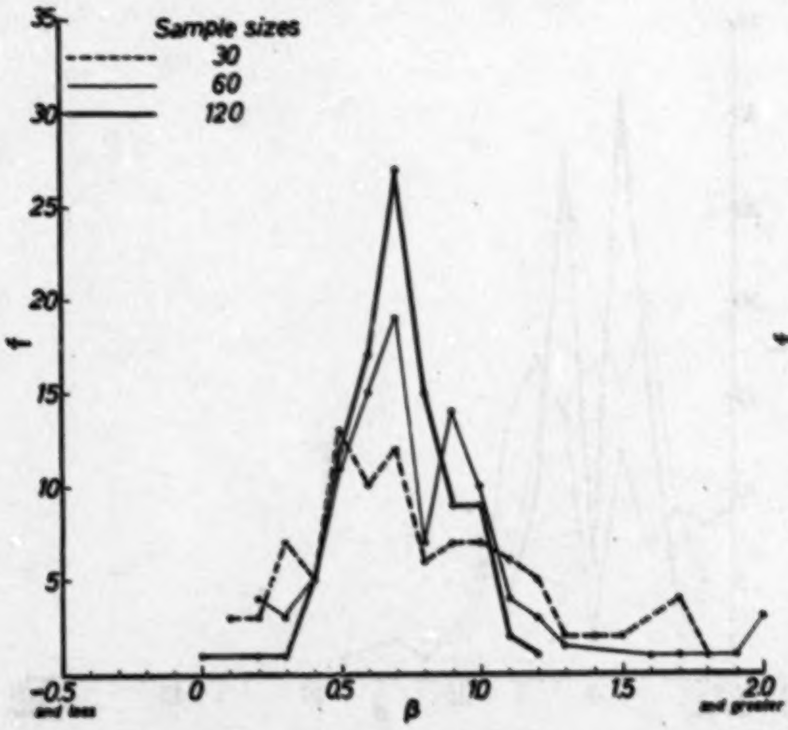


Figure 30i. Item 9, internal criterion, β =.636

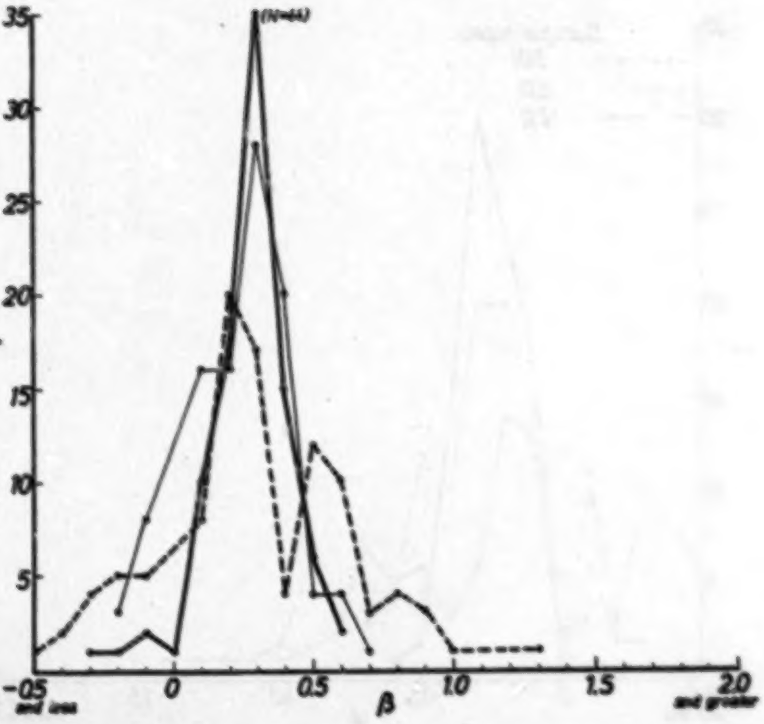


Figure 30j. Item 1, external criterion, β =.205

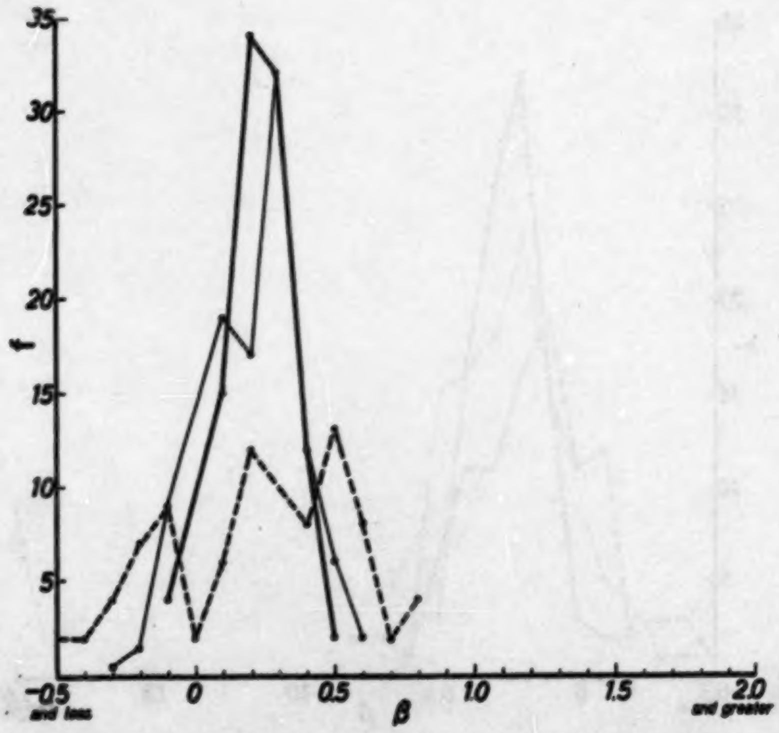


Figure 30k. Item 2, external criterion, β =.205

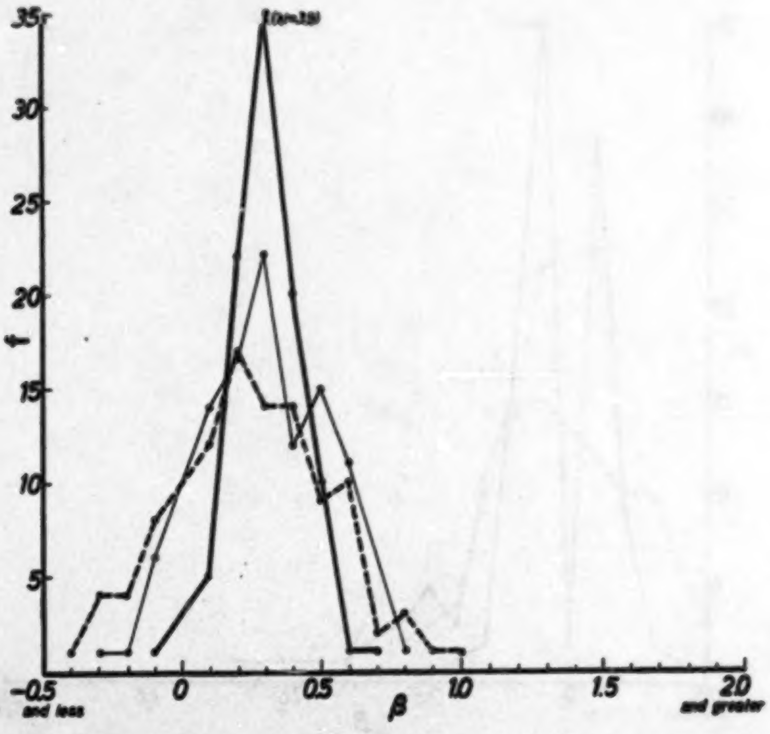


Figure 30l. Item 3, external criterion, β =.234

Figure 30 continued.

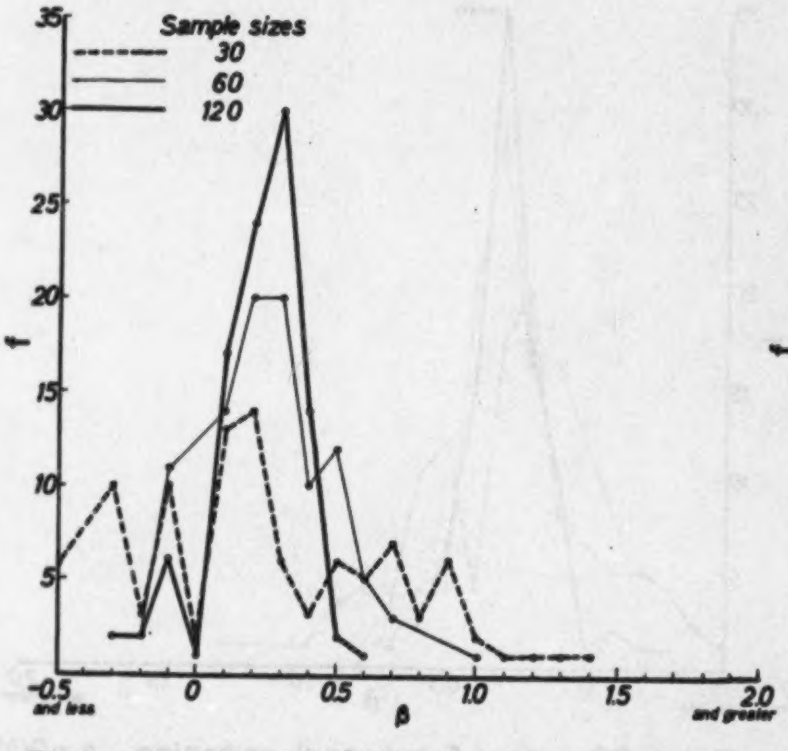


Figure 30m. Item 4, external criterion, β = .202

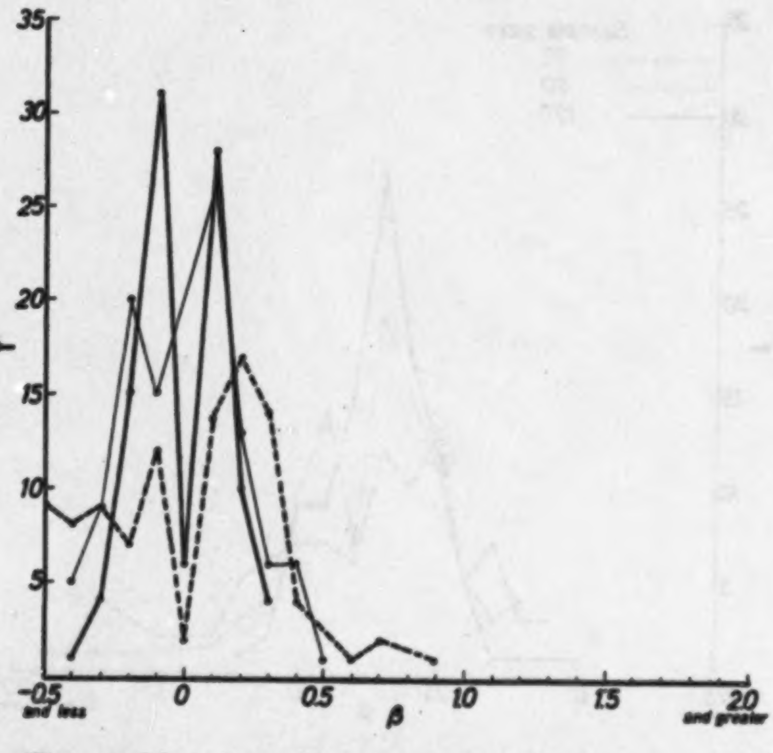


Figure 30n. Item 5, external criterion, β = -.013

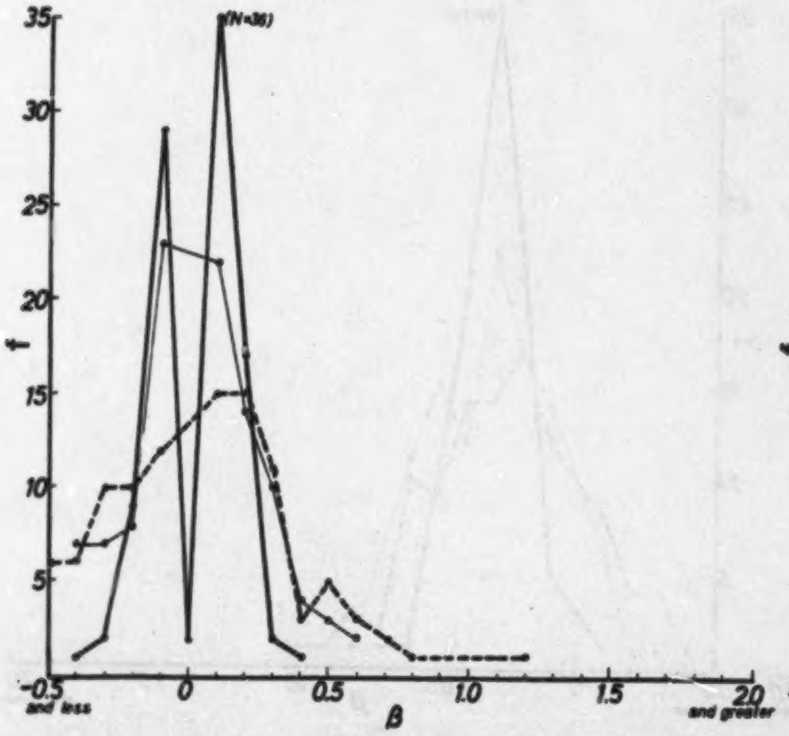


Figure 30o. Item 6, external criterion, β = .067

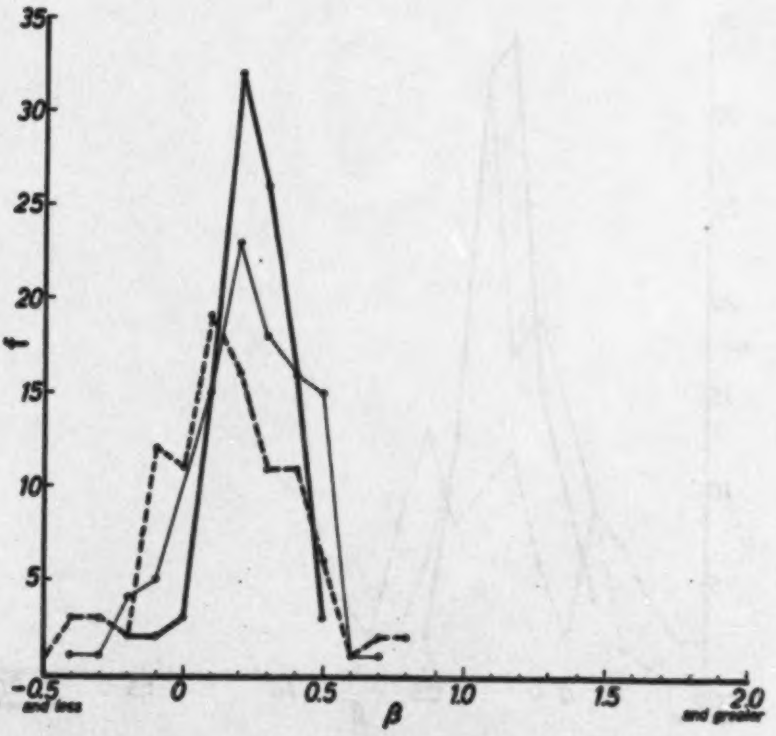


Figure 30p. Item 7, external criterion, β = .192

Figure 30 continued.

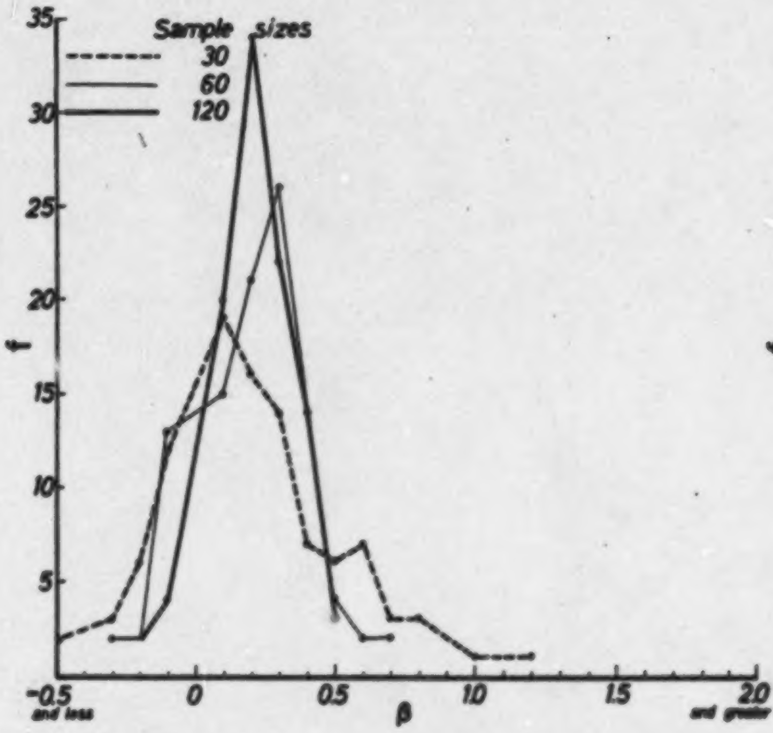


Figure 30q. Item 8, external criterion, $\beta = .144$

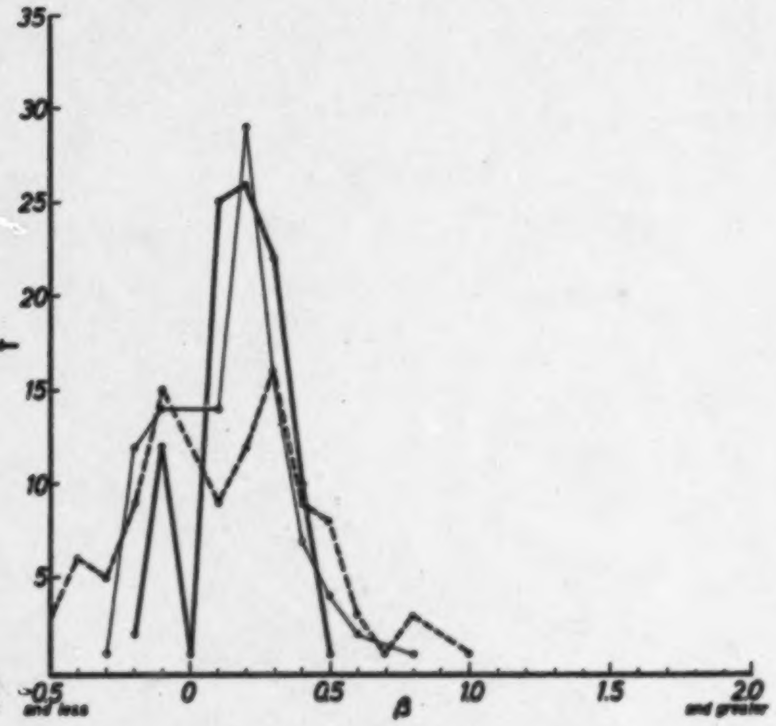
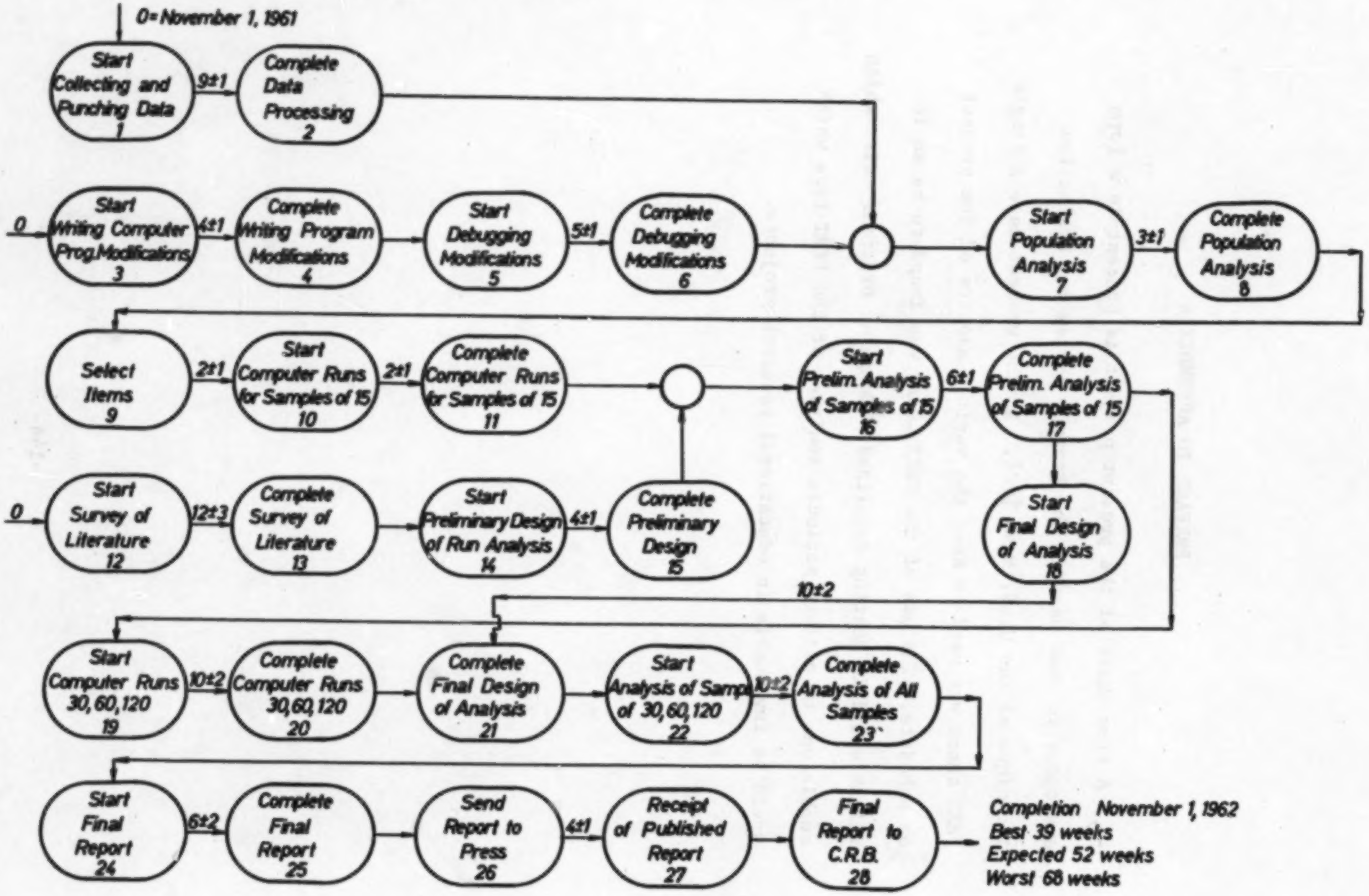


Figure 30r. Item 9, external criterion, $\beta = .149$

PREFACE TO APPENDIX B

A flow chart of the present project is present in a form analogous to that used in the Program Review and Evaluation Technique of the U. S. Navy [42]. In the present case a single PERT chart was used to keep the various aspects of the project on schedule. The use of the PERT chart was found to be an invaluable aid in meeting deadlines. Computer programs exist which enable one to perform periodic analyses of the PERT type which would be invaluable in educational research projects.

Figure 31. Project pert chart.


0 = November 1, 1961
Start Collecting and Punching Data 1
9±1
Complete Data Processing 2
0
Start Writing Computer Prog. Modifications 3
4±1
Complete Writing Program Modifications 4
Start Debugging Modifications 5
5±1
Complete Debugging Modifications 6
Start Population Analysis 7
3±1
Complete Population Analysis 8
Select Items 9
2±1
Start Computer Runs for Samples of 15 10
2±1
Complete Computer Runs for Samples of 15 11
Start Prelim. Analysis of Samples of 15 16
6±1
Complete Prelim. Analysis of Samples of 15 17
0
Start Survey of Literature 12
12±3
Complete Survey of Literature 13
Start Preliminary Design of Run Analysis 14
4±1
Complete Preliminary Design 15
Start Final Design of Analysis 18
10±2
Start Computer Runs 30, 60, 120 19
10±2
Complete Computer Runs 30, 60, 120 20
Complete Final Design of Analysis 21
Start Analysis of Samp of 30, 60, 120 22
10±2
Complete Analysis of All Samples 23
Start Final Report 24
6±2
Complete Final Report 25
Send Report to Press 26
4±1
Receipt of Published Report 27
Final Report to C.R.B. 28
Completion November 1, 1962
Best 39 weeks
Expected 52 weeks
Worst 68 weeks